7111

TESTS AND MEASUREMENTS IN INDUSTRIAL EDUCATION

TESTS AND MEASUREMENTS IN INDUSTRIAL EDUCATION

BY

LOUIS V. NEWKIRK, PH.D.

Director, Division of Industrial Arts, Chicago Board of Education

AND

HARRY A. GREENE, PH.D.

Professor of Education and Director of Bureau of Educational Research and Service, State University of Iowa

NEW YORK

JOHN WILEY & SONS, INC.

LONDON: CHAPMAN & HALL, LIMITED

FOURTH PRINTING, AUGUST, 1949

PRINTED IN THE U. S. A.

PREFACE

The growth of interest in tests and measurements in industrial education has been rapid in recent years. The last decade has produced numerous significant investigations in the curricular aspects of these special subjects, on which have been built newer and better materials and methods of teaching. A renewed interest in the possibilities of measurement of special aptitudes and achievement in the industrial education fields naturally parallels this type of development.

This book is planned to fit into this program. It is designed to bring to the attention of the shop teacher, and to students in training for this type of work, a simple and practical discussion of the essential principles of educational measurements as applied to the teaching of shop and drawing courses. It is based upon a considerable number of years of experience on our part in the teaching of courses in educational measurements and in methods in the industrial arts fields. In addition to these major functions, this book is planned to stimulate a renewed interest in the more adequate evaluation of student achievement by teachers of industrial education who have already had some experience with the work. It brings together and evaluates many of the more important contributions to measurements in industrial arts and industrial education. We earnestly hope that it may also serve to stimulate further interest and work along these lines.

In presenting this material we recognize the difficulty of covering in an adequate manner the many difficult problems. Throughout the book, the aim has been to emphasize the practical rather than the theoretical. It is not planned to displace general treatises on measurements or statistics. On the other hand, it is hoped that the straightforward presentation of the problems of measurement in this subject may eliminate the necessity for technical training in measurements and statistics in order for the student or teacher to use this book effectively.

We wish to acknowledge our great indebtedness to the many classroom teachers and supervisors who have contributed directly and indirectly to the materials presented in this discussion. The kindness of authors and publishers who have given permission for the reproduction of many selected portions of their work and publications is

likewise gratefully acknowledged. We are also indebted to Professor Arthur B. Mays of the University of Illinois, who gave valuable editorial criticisms; to Professor A. H. Edgerton of the University of Wisconsin for encouragement and valuable suggestions; to President Butler Laughlin of the Chicago Normal College for editorial suggestions; and to Professor Frank X. Henke of the Chicago Normal College for illustrative drawings.

Chicago, Ill.,
May, 1935.

L. V. Newkirk
H. A. Greene

CONTENTS

LIST OF TABLES

TESTS AND MEASUREMENTS IN INDUSTRIAL EDUCATION

CHAPTER I

INTRODUCTION

1. Significance of Measurement in Industrial Education.

Measurement in its various forms and phases appears to be recognized as an integral part of good classroom procedure. In no instructional field is there greater need for the application of the principles of educational measurement than in industrial education. Industrial education [1] teachers and supervisors need reliable measuring instruments in order to give more adequate educational guidance, to evaluate personality traits, to motivate learning, to study the effectiveness of teaching materials and methods, to measure pupil progress more accurately through the establishment of more definite standards of performance and through the diagnosis of pupil difficulties. The use of tests for such purposes in other fields is a well-established practice. The fundamental principles of scientific test construction and interpretation may be applied to the measurement problems of the shop and the drafting room when modified in the light of special needs. It is the purpose of this book to explain and illustrate many of the applications and modifications of recognized principles of measurement to these specific fields.

The marked increase in the interest in educational measurements on the part of teachers of industrial education is not surprising. It is to be expected of a group of teachers who have had to face the many problems of a new and growing unit of instruction. In many ways the teachers of industrial education are most fortunate. They are working

[1] Throughout this text the term *industrial education* is used to include the general courses of the secondary school variously known as manual training, manual arts, industrial arts, and industrial arts education, and the vocational work of the continuation school, trade school, and evening schools. The fundamental measurement problems in all these courses are similar, although the objectives of the courses vary from cultural to strictly vocational.

in a new and growing field of instruction which rapidly is becoming organized in the light of modern educational objectives. They have the advantages of all the methods and techniques that have been developed for measurement in other fields. They are in a position to utilize the good and discard the worthless results of earlier efforts. A large number of accepted principles and practices for use in constructing and interpreting measuring instruments are now available for the evaluation of the products of teaching. From the standpoint of professional qualifications and classroom efficiency it is the industrial education teacher's business to understand these well-established principles and their special application and use in their own fields of instruction.

2. Teachers' Marks in Industrial Education.

The earlier studies of teachers' marks revealed the fact that such measures were entirely unsuited for the evaluation of pupil achievement since they were extremely subjective and quite unreliable. However, these earlier studies confined themselves mainly to teachers' marks used in the rating of accomplishment as it is revealed on the written page. It is true that many of the teachers' marks in industrial education are given on this same basis, but there is also the matter of rating actual projects and drawings from the shop and drafting room as well as the manipulative skills. The absence of precise information on the exact subjectivity and unreliability of such marks in the industrial education fields prompted the authors to carry on a series of investigations in this field. The results are presented here as further evidence of the need for improved methods of measuring the accomplishment of students in these subjects.

Three samples from each of the fields of woodworking, drawing, and sheet metal were selected for study. The woodworking projects consisted of one gray wren-house, one red wren-house, and one rolling pin. The drawings were simple inked drawings, and were known as numbers 1, 6, and 7. (See Fig. 1.) The sheet-metal projects consisted of three funnels which were numbered 1, 2, and 3.

A group of experienced industrial education teachers coöperated in rating these projects. The marking was done through individual or group conferences with the teachers and in accordance with the following instructions:

1. Pay no attention to the name of the maker of the projects but rate them entirely on the basis of what you would consider perfect.
2. Rate the projects on a scale of 100, giving 100 for perfect, 50 for half perfect, etc.
3. List the factors that you took into consideration in rating the projects.

The teachers were asked to rate the project only on the basis of what they considered perfect, and to give no consideration to grade standards for such a project. The factor of grades was avoided, since it was the main purpose here to discover how much variability there is among

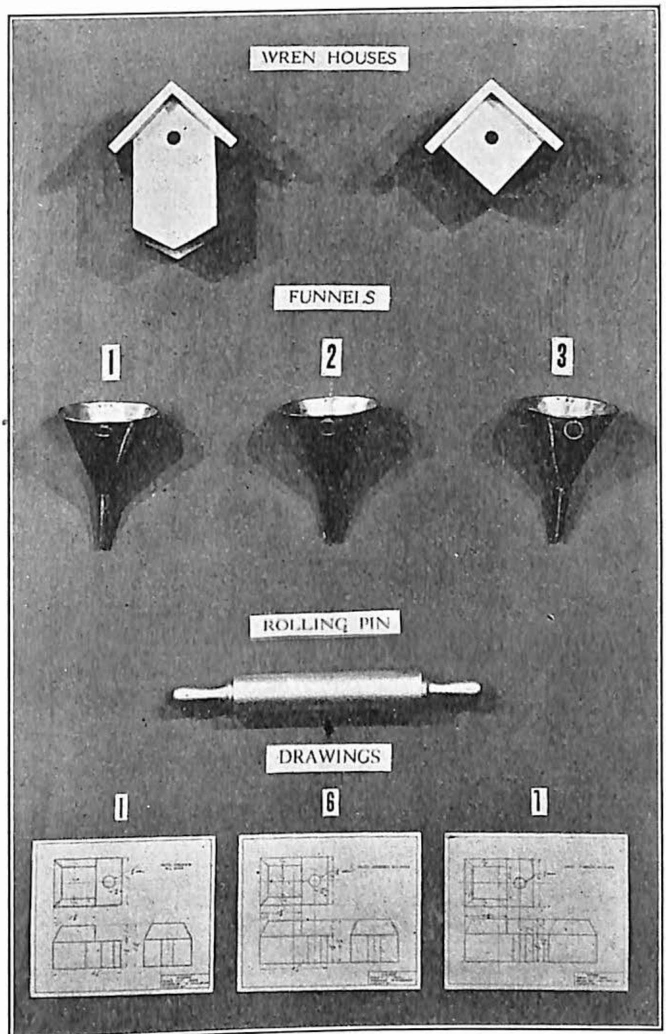


FIG. 1.—Samples for rating projects.

teachers in their concept of what is good workmanship. The marks and rating factors obtained on beginning woodworking, drawing, and sheet-metal projects are given in Tables 1 to 6 inclusive.

Results of Rating Woodwork Samples. Table 1 shows a range of 40 to 81 for the ratings of the red wren-house, 70 to 96 for the gray

TABLE 1

RATINGS ASSIGNED WOODWORK SAMPLES

(39 teachers)

Red Wren-house			Gray Wren-house			Rolling Pin		
Rating	Frequency	Mark	Rating	Frequency	Mark	Rating	Frequency	Mark
81	1	C	96	1	A	98	1	A
80	3	C	95	5	A	96	1	A
78	4	D	94	1	A	95	6	A
75	10	D	92	3	B	94	2	A
74	1	D	90	11	B	93	2	A
73	1	D	87	1	B	92	3	B
70	3	F	85	12	C	90	14	B
65	2	F	80	2	C	85	2	C
61	2	F	75	2	D	83	1	C
55	1	F	70	1	D	80	1	C
53	1	F				78	1	D
51	1	F				75	3	D
50	6	F						
40	3	F						

wren-house, and 75 to 98 for the rolling pin. There is also a tendency for the ratings to bunch at certain points on the scale. The letter grades show in a rough way that pupils with projects of the same basic quality might be assigned almost any of the variable passing marks, depending upon which shop teacher rated them. It must be remembered, however, that these results are not in all respects comparable to the actual situation in the shop or at the drawing table. In either of these situations the teacher almost certainly would grade on the class average. Furthermore, it would be necessary to take into consideration the physiological development of the pupils, their intelligence quotients, and quite likely their mechanical aptitudes. There would also be the factor of the pupil's personality and class attitude which might affect the teacher's judgment.

The data given in Table 2 show clearly that there is a wide variation in the factors mentioned in scoring the same projects. There is a distinct preference, however, for rating factors which group themselves under the following headings: results of tool operations, finish, design, fasteners, and utility. It appears, therefore, that much of the variation in the rating of these projects was due to the varying amounts of emphasis placed on these factors by the teachers themselves.

Rating Drawing Samples. The ratings of the drawings (Table 3) show more variation and less tendency to cluster than is found in the

TABLE 2

MAJOR FACTORS CONSIDERED BY JUDGES IN RATING THE THREE WOODWORKING PROJECTS

(39 teachers)

Rating Factors	Frequency
Finish	31
Joints	30
Proportion	23
Squareness	22
Nailing	19
Utilitarian	13
Commercial standard of workmanship	13
Design	10
Sanding	8
Fitting	7
Gluing	5
Dimensions	4
Choice of materials	4
Planing	4
Shape	4
Accuracy	3

TABLE 3

RATINGS ASSIGNED THREE BEGINNING DRAWING PROJECTS

(27 teachers)

Sample 1			Sample 6			Sample 7		
Rating	Frequency	Mark	Rating	Frequency	Mark	Rating	Frequency	Mark
94	1	A	88	1	B	99	1	A
90	2	B	87	1	B	98	1	A
85	3	C	86	1	C	97	1	A
84	1	C	85	2	C	95	4	A
82	1	C	82	1	C	93	2	A
80	7	C	80	2	C	90	8	B
75	4	D	78	1	D	85	6	C
70	3	D	75	3	D	80	3	C
65	1	F	70	6	F	75	1	D
60	1	F	65	3	F			
40	1	F	60	1	F			
25	1	F	55	2	F			
			50	1	F			
			30	1	F			

ratings of the woodworking projects. The two poorer drawings (Samples 1 and 6) show the most variation and the best drawing the least, although this drawing (sample number 7) has a variation equal to the range of all of the passing marks.

The frequencies given in Table 4 indicate that the drawing teachers considered the factors of lettering, figures, and lines most often, but they also took into consideration neatness, dimensions, erasures, arrowheads, and accuracy with fair consistency. On the whole, the

TABLE 4

MAJOR FACTORS CONSIDERED BY JUDGES IN RATING THE THREE DRAWINGS

Rating Factors	Frequency
Lettering and figures	62
Lines	62
Neatness	27
Dimensions	22
Erasures	18
Arrowheads	16
Accuracy	13
Cleanliness	8
French curves	8
Placement	8
Completeness	7
Spacing	5
Projection	5
Joints	4
General appearance	4

drawing teachers showed a slightly greater variation than the woodworking teachers in their ratings.

Rating Sheet-Metal Projects. Differences quite typical of all such ratings are found for the three sheet-metal projects used in the study (Table 5). It is apparent from Table 6 that the quality of the soldering was the factor considered most critically by these teachers.

Summary of Ratings. The results of this study indicate that shop and drawing teachers are highly unreliable in their ratings of the same group of projects or drawings. Furthermore, they do not agree on the relative importance of the factors to be considered in making such ratings. There are probably not enough cases to warrant the generalization that teachers of shop work and drawing are any more or less subjective in their markings than are teachers of the academic subjects. Yet the study does indicate that there is sufficient variation in

TABLE 5

RATINGS ASSIGNED THREE SHEET-METAL PROJECTS
(12 teachers)

Sample 1			Sample 2			Sample 3		
Rating	Frequency	Mark	Rating	Frequency	Mark	Rating	Frequency	Mark
90	1	B	95	2	A	96	1	A
87	1	B	92	1	B	90	3	B
85	2	C	90	1	B	85	1	C
80	1	C	82	1	C	80	4	C
77	1	D	80	1	C	78	1	D
75	1	D	75	5	D	75	1	D
70	1	F	70	1	F	60	1	F
60	1	F						
55	1	F						
50	1	F						
40	1	F						

the estimation of quality in these specimens to introduce serious errors in measurement based on such a procedure. Table 7 summarizes the range in ratings and the corresponding letter marks for the nine projects included in the study.

TABLE 6

MAJOR FACTORS CONSIDERED BY JUDGES IN RATING THE THREE SHEET-METAL PROJECTS

Rating Factors	Frequency
Soldering	33
Proportion	13
Seams	13
Roundness	12
Shape	11
Wiring	9
Neatness	8
Accuracy	6
Forming	6
Design	6
Roughness	5
Joints	4
Curve	2
Crimping	2

TABLE 7

SUMMARY OF RATINGS

Project	Range of Ratings	Corresponding Letter Marks
Woodwork	41; 26: 23	C-F; A-D; A-D
Drawing	69; 58; 24	A-F; B-F; A-D
Sheet metal	50; 25; 38	B-F; A-F; A-F

3. Need for a Knowledge of Measurements in Industrial Education.

The rating of shop projects and drawings is difficult; it requires a complex fusing of judgments based on a group of variable factors. Yet, psychologically, the rating of shop projects and drawings is little different from the rating of an English theme or a paper in mathematics. In English the factors to be considered may be spelling, sentence structure, paragraphing, punctuation, etc.; and in shop subjects the judgment may be based on such factors as tool processes, design, utility, finish, and fasteners. Teachers vary greatly in their concept of what constitutes perfection in a project or drawing. The same project or drawing looks different to different individuals, and quite probably to the same individual under different circumstances.

Marks assigned by teachers of shop work and drawing evidently are subject to the same types of errors as enter into all estimates of achievement. The magnitude of the error is sufficient also to warrant the conclusion that industrial education teachers need objective measurements of the results of their teaching just as much as instructors in any other field. The fact that numerous tests capable of securing valid and reliable measures of subject-matter and tool skills in this field have been constructed is proof that the fundamental principles of test construction can be successfully applied to the measurement of the results of teaching in shop work and drawing. The problem here is not to develop new principles of testing, but rather to modify, apply, and illustrate in the industrial arts field those procedures which have been generally found to be sound and fundamental in other subjects.

In addition to securing accurate measures of the results of teaching, there is the equally important need on the part of teachers of understanding the student better, in order that they may be in a position to give him the proper guidance in the selection of special lines of training. To accomplish this, industrial education teachers must know the individual levels of general intelligence, special apti-

tudes, informational background, interests, appreciations, and emotional traits and attitudes of their students.

SUMMARY EXERCISES FOR DISCUSSION

1. What specific factors appear to make objective measurement in industrial education quite difficult?
2. Rate a project in woodworking, drawing, or metal working, on a percentage basis, and record the characteristics of each project which influenced you most in assigning the marks.
3. If you have access to a class, have each student mark independently a project in each of the above fields, and compare the marks as to variability, following the procedure shown in Table 1.
4. Tabulate the characteristics of each project that were mentioned by the stuents as being considered in marking the project.
5. In your judgment, what factors largely account for the wide variation in marks of achievement assigned to products of a similar quality?
6. Suggest a number of devices which would seem to have possibilities for increasing accuracy in the assignment of marks to shop projects.

SELECTED REFERENCES

HUNTER, WILLIAM L., "Objective Tests in Shop Courses," *Industrial Education Magazine,* Vol. 29: 433-39, No. 12, June, 1928.

KELLY, F. J., *Teachers' Marks,* Teachers College Contribution to Education, No. 66, p. 11, Columbia University, 1914.

LEAVITT, F. M., "Standardized Measurements in the Field of Industrial Arts," *Industrial Arts Magazine,* Vol. 8: 132, April, 1919.

MANSPERGER, D. E., "Testing the Industrial Arts in Junior and Senior High School," *Industrial Arts Magazine,* Vol. 18: 49, 1929.

MEYER, MAX, "The Grading of Students," Science (N. S.), Vol. 28: 243-252.

NASH, HARRY B., and VAN DUZEE, ROY R., "The Standard Test in Industrial Arts," *Industrial Arts Magazine,* Vol. 19: 125-29, No. 4, April, 1930.

NEWKIRK, L. V., "Reliability of Shop Teachers' Marks in Rating Shop Projects and Drawing," *Industrial Arts and Vocational Education Magazine,* Vol. 20: 123, April, 1931.

SMITH, HOMER J., "Objective Measurement in Industrial Education," *Industrial Education Magazine,* Vol. 31: 331-336, No. 9, March, 1930.

STARCH, DANIEL, and ELLIOTT, EDWARD C., *School Review,* Vol. 20: 442-57; 21: 254-59; 26: 676-81.

SWOPE, AMMON, "How to Construct Objective Tests in Industrial Subjects," *Industrial Education Magazine,* Vol. 30: 7-9, No. 1, July, 1928.

CHAPTER II

TYPES OF EDUCATIONAL TESTS

4. Essay-Type Tests.

The two general classifications of educational tests in common usage are objective and essay-type tests. Objective tests are so constructed that they can be scored without any guessing or subjective judgment on the part of the user. In the traditional or essay test a number of questions are made out covering the material to be tested in a general way with statements similar to the following:

1. Name ten common cabinet woods.
2. How are the grades of sandpaper indicated?
3. What is the difference between spindle turning and face-plate turning?
4. What is varnish?
5. What is the princple of the internal-combustion engine?

The average teacher using the essay-type examination makes up five or ten questions on the subject being tested (drawing, woodwork, sheet metal, auto mechanics) and then allows the pupils thirty to fifty minutes to answer them. The directions for administering such a test usually consist in a statement reminding the pupils to write their names on each sheet before handing in the test.

The scoring of the essay-type examination presents a real problem, some phases of which were introduced in the preceding chapter. The teacher's principal object in giving the test is usually to secure an estimate of the pupil's mastery and retention of the informational content of the course. In correcting an essay-type examination, factors appear which influence the teacher's judgment but which have little to do with the actual evaluation of the student's knowledge of the subject. Some of these factors are English, including spelling, sentence structure, paragraphing, composition; mechanical features of the examination such as neatness, legibility, use of pen or pencil, use of one or both sides of the paper, kind and size of paper used; the quantity written; the sampling of the subjects represented by the questions; the teacher's attitude toward the pupil, or the pupil's attitude

toward the teacher. The final mark is influenced by unknown combinations of these factors. This means that the mark on the test is an entirely inadequate expression of any one factor, and hence is an unreliable measure of the entire field covered by the test. Thus the essay test at best can furnish only the roughest measure of achievement.

It is frequently argued by those defending the essay-type test that it gives the student valuable training in the mechanics of writing, spelling, thought organization, and expression. If this were actually accomplished, the argument would be sound, but even an unbiased observer of students engaged in writing essay-type examinations must admit that the rush and strain of getting the words down on the examination paper leaves very little opportunity for the training in thought organization and expression which it should give. It seems safe to conclude, therefore, that if a teacher desires to measure a pupil's ability to spell, write, and express himself, he should use tests designed for that purpose and not confuse the issues.

5. Objective Tests.

Properly constructed objective-test exercises are not influenced appreciably by the conflicting factors which appear to invalidate measurement based on the essay-type question. Objective exercises are marked by two important and related features. These are (1) brevity of pupil response, and (2) absence of personal judgment in scoring the test exercises. These features of the objective exercise make it equally suitable for use in the teacher-made informal examination, and in the more carefully constructed standardized test.

Objective exercises are stated in such forms that the pupil is able to indicate his understanding by the briefest and simplest of physical responses, usually consisting of underlining or encircling a single word or phrase. Because of this brevity of pupil response, many more exercises may be submitted to the pupil, thus providing a more complete sampling or coverage of the subject-matter. The quality of the answers need not be evaluated by the teacher but may be scored as right or wrong by comparison with an answer key. The use of the objective form of the test exercise thus makes it possible for different teachers to score the same test papers and secure identical results. A test exercise which is perfectly objective may be scored repeatedly at widely separated intervals and by different individuals without significant variation in results. Such accuracy in grading test exercises can be obtained only when the exercises are constructed in accordance with certain rather well-known specifications.

6. Tests and Scales.

Measuring instruments are roughly divided into *tests* and *scales.* This distinction is of some value, but at times it is confusing because some tests resemble scales or contain certain features of scales as an essential part of their construction. Generally speaking, *a test is a measuring instrument used for the evaluation of any knowledge, quality, or ability.* It may measure degree of achievement, mental ability, aptitude, or character traits. It may be made up of items of uniform difficulty, or it may be composed of a series of items of uniformly increasing difficulty or value. In the former case it is a *rate* test; in the latter, it is a *power* test. The process of determining the difficulty or value of test items is called *scaling.* The use of this term possibly accounts for much of the confusion concerning tests and scales.

A measuring instrument is a *scale* to the extent that it ranks accomplishment directly in terms of systematic levels, grades, or ages. An instrument which is made up of scaled items (items of systematically increasing difficulty) and which expresses its results in terms of the number of such items responded to correctly is still a test. Such a test is made quite often by the selection of items of known value from a scale. For example, a spelling test comprising words of gradually increasing difficulty could be made from the *Simmons-Bixler High School Spelling Scale* (see brief description of this scale, page 99) by selecting the test words from columns in which a uniformly decreasing percentage of pupils spell the words correctly. This test might be treated as a scale if the scores on it were expressed in terms of the scale value of the last word spelled correctly. If the results were expressed in terms of the total number of these words spelled correctly it would be considered a test.

Most modern tests are really hybrids resulting from the cross-breeding of these two forms. That is, they are made up of scaled items, but the resulting scores representing accomplishments are expressed in terms of the number of items responded to correctly. The specimen test shown on page 140-144 is an illustration of this type.

In the evaluation of accomplishment in industrial education the quality scale has numerous uses. In general, scales of this type consist in a series of specimens of the particular quality under consideration arranged in ascending order of merit from very poor (or zero quality) to very high quality. The accomplishment of the individual student is expressed in terms of the value of the specimen on the scale most nearly matching his own product. Obviously, the use of such quality scales introduces a considerable amount of subjectivity into the meas-

urement, since the teacher's judgment is necessarily involved in assigning the quality rating. Such scales are used for the rating of handwriting, free-hand lettering, drawing, electrical splicing, soldering, wood-boring, riveting, forging, finishing, and many other products. The techniques used in the construction and use of these scales are discussed and illustrated in Chapter XII.

7. Standardized and Informal Objective Tests.

Objective measuring instruments are further designated as standardized tests and teacher-made tests and scales. Both types are useful in measuring achievement in industrial education. A test is standardized (1) if it is composed of exercises that have been selected in the light of usual teaching practice and evaluated as to innate difficulty, and (2) if it is accompanied by norms or standards permitting the interpretation of results in levels of accomplishment. Standardized tests are of value in comparing the accomplishment of a class with general standards and in comparing groups in different schools in the same system. Teacher-made or informal objective tests are similar to standardized tests except that the test items are selected directly from the content of the course of study. Usually the items in such tests more closely parallel the material taught but are less carefully formulated and evaluated than standardized tests. Generally, too, no norms are available, but useful levels of accomplishment may be developed from year to year by recording the scores each time the test is given. Teacher-made objective tests are extremely useful in measuring achievement and diagnosing instruction in the shop.

From the standpoint of their administration, standardized and teacher-made tests may be classified as written, oral, and performance. Psychologically there is little difference, because, after all, they are all performance tests. It has been found advantageous in testing different types of industrial education achievement to have the pupils write the responses to some items, respond orally to others, and in many cases express their knowledge by modifying material through the use of tools and machines. The fundamental thing to note here is that in order to measure scientifically it is necessary to secure a response which can be rated objectively and compared with the same response made by others.

8. Classification of Educational Tests.

Educational measuring devices may be classified according to their use and characteristics into achievement, diagnostic, prognostic, and intelligence tests. Each of these types of instruments is useful in measuring class and individual accomplishments, and in revealing the

general and special capacities of students in the industrial education subjects.

Achievement Tests. Achievement tests measure abilities or products acquired from the school or other types of educational experience of the pupil. Such tests may be standardized, or they may be informal examinations made by the teacher. Considerable attention is given in this book to the problems arising out of the construction, use, and evaluation of achievement tests.

Diagnostic Tests. Diagnosis is really one of the major underlying purposes of all achievement testing. In fact, it may be said that all general achievement tests are diagnostic to a degree. Most achievement tests, however, fail to furnish adequate diagnostic information because of the large number of skills they cover and because of the difficulty of securing a sufficiently detailed interpretation of the results. Diagnostic tests are specially constructed achievement tests designed to discover the exact identity and location of the pupils' strengths and weaknesses in subject-matter mastery. The development and use of such tests mean, of course, that the subject-matter itself has been analyzed to the point that the basic or underlying skills are clearly identified. It is fairly safe to assume that subject-matter fields in which detailed diagnostic tests are not available have not yet been subjected to this type of analysis.

Tests of this diagnostic or analytical character, were they available, would be most useful to industrial education teachers in discovering what is already known by the pupil and thus indirectly in finding what remains to be mastered. This is really an inventory use of the tests. Genuine diagnostic tests have been slow to appear in industrial education subjects. The *Newkirk-Stoddard Home Mechanics Test* [1] (see page 115 for extracts from this test), though not strictly diagnostic, furnishes a useful analysis of instruction in home mechanics. It may be used also to determine how well a school is teaching the outstanding home mechanics jobs or what jobs the individual pupils are best acquainted with. Hunter's *Shop Tests* [2] are further illustrations of tests with some diagnostic value. For example, in this series of measures on woodwork there are tests on tools, fastenings, trade names, reading rules, wood finishing, and others.

Prognostic Tests. One of the very significant features of modern measurement in education is its emphasis on prediction. Tests of gen-

[1] Newkirk, L. V., and Stoddard, George D., *Newkirk-Stoddard Home Mechanics Test*, Bureau of Educational Research and Service, State University of Iowa, Iowa City, Iowa, 1928.

[2] Hunter, Wm. H., *Shop Tests*, The Manual Arts Press, Peoria, Illinois.

eral mental ability are useful to the extent that they predict a pupil's general level of accomplishment. Prognostic tests are measures of specialized aspects of intelligence. The purpose of such tests is to provide the basis for accurate prediction of future achievement in specialized fields on the basis of present performance on some fundamental underlying elements of the subject. Prognostic tests are designed to measure specific abilities underlying achievement in a particular subject-matter field rather than the achievement itself.

Aptitude or prognostic tests in industrial education should be most useful in determining the probability of success of a student in such subjects as drawing, machine shop, carpentry, bricklaying, cabinet making, or in any other special field. Tests of mechanical ability have value in predicting probable future success in industrial education subjects.

Intelligence Tests. There are many definitions of intelligence, and many different ways and means of measuring it. In general, intelligence is the capacity of the individual to adapt himself to novel situations. It is the power of the individual to learn. In actual practice, intelligence is usually measured in terms of the extent to which the individual has applied this power in the acquisition of information and skills in a number of specific and mainly unrelated fields. In a sense, general mental ability is like a cable composed of many strands and fibers of varying size and quality, each representing some particular phase of ability. The intelligence test is merely a device for taking a cross-section of this cable. If the measuring device reveals the large and important strands of the table it is a valid instrument.

SUMMARY

The two general classifications of educational tests in common usage are objective and essay-type tests. The objective test may be scored without the subjective judgment of the teacher. The grading of the essay type of examination presents a real problem and is influenced by the subjective judgment of the teacher. A test exercise which is perfectly objective may be scored repeatedly at widely separated intervals without significant variation in results.

Measuring instruments are roughly divided into tests and scales. A test is a measuring instrument used for the evaluation of any knowledge, quality, or ability. A scale is a measuring instrument that ranks accomplishment directly in terms of systematic levels, grades, or ages. Objective measuring instruments are further designated as standardized and teacher-made tests and scales. A test is standardized when it is composed of exercises that have been selected in the light of usual

teaching practice, evaluated as to innate difficulty, and is accompanied by norms or standards permitting the interpretation of results in levels of accomplishment.

Achievement tests measure abilities or products acquired from the school or other types of educational experience of the pupil. Diagnostic tests are specially constructed achievement tests designed to discover the exact identity and location of the pupil's strengths and weaknesses in subject-matter mastery. Prognostic tests may be thought of as measures of specialized aspects of intelligence. Intelligence or general mental ability may be described as the power the individual has to adapt himself to novel situations. Intelligence tests are classified as group and individual, depending on the method of administration they employ.

SUMMARY EXERCISES FOR DISCUSSION

1. What special features distinguish the objective test from the essay-type test?
2. Enumerate as many as possible of the special factors which distinguish standardized tests from informal objective tests.
3. What does the process of standardization of a test imply?
4. Illustrate the different types of educational tests, using materials from the industrial arts field.
5. In what specific ways are prognostic tests different from tests of general mental ability?
6. What distinguishes a test from a scale?
7. What qualities distinguish a rate test from a power test?
8. Suggest specific ways in which quality scales may be particularly useful in industrial arts classes.
9. What types of measuring instruments seem to have the greatest possibilities of practical value in industrial education?
10. Do you think it will ever be possible to develop genuinely diagnostic tests in this field of instruction? Why?

SELECTED REFERENCES

Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests.* New York: Longmans, Green and Company, 1935.

Kelley, T. L., *Interpretation of Educational Measurements.* Yonkers, New York: World Book Company, 1927.

Lang, A. R., *Modern Methods in Written Examinations.* Boston: Houghton Mifflin Company, 1930.

Monroe, W. S., *The Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923.

Odell, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

Odell, C. W., *Traditional Examinations and New-Type Tests.* New York: The Century Company, 1928.

RUCH, G. M., *The Objective, or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929.

RUCH, G. M., and STODDARD, G. D., *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927.

SMITH, H. L., and WRIGHT, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

SYMONDS, P. M., *Measurement in Secondary Education.* New York: The Macmillan Company, 1927.

WILSON, G. M., and HOKE, K. J., *How to Measure* (Revised). New York: The Macmillan Company, 1928.

WOOD, BEN D., *Measurement in Higher Education.* Yonkers, New York: World Book Company, 1923.

CHAPTER III

USES OF TESTS IN CLASSROOM AND SHOP

9. Tests as Related to Instruction.

The uses of educational tests for administrative, supervisory, research, and survey purposes, important as they are, do not represent their most vital and important functions. In the past so much emphasis has been given to these particular uses that the teacher often lost sight of their real utility in the solution of his individual instructional problems. The recent development of reliable, valid, and highly detailed measuring instruments designed to parallel closely the subject-matter content taught by the teacher has caused him to shift his point of view. He realizes now that tests are most important supplements to other instructional material, and that without them he can scarcely hope to work at his highest level of instructional efficiency. He notes that modern tests are usually well made, detailed, comprehensive, and analytical. He sees that with this type of instrument available it is possible for him to test as he teaches; to chart his instructional course from accurate and objective observations. Modern tests give the teacher a chance to discover where emphasis should be placed, and to determine when a satisfactory level of control has been attained.

The busy classroom teacher can hardly be expected to construct tests which will possess all the merits of a carefully constructed standard test. This assumes a breadth of knowledge of the subject-matter and a training in the technique of test construction which most teachers do not have. Even if there were perfect subject-matter mastery and a thorough knowledge of the making of tests, it is doubtful if the typical classroom teacher should be expected to spend his time in this way when, in most fields, other, better-made, and more economical materials are available. Yet, the teacher should certainly not be forced to depend upon his general observation of his pupils for his information concerning their strengths and weaknesses.

There are times and conditions, however, in which the informal objective or teacher-made test is very useful. The teacher-made objective-type shop test serves its most useful function in measuring the

relative achievement of the individual members of a class. Quite often standardized tests contain items that are not taught in the course. Frequently they do not contain items which are taught. Teacher-made tests can be designed to fit the specific needs of the shop teacher's own course. Teacher-made achievement tests can be used for diagnosis of special difficulties, for the motivation of learning, for the measurement of accomplishment and assigning shop marks; but since they do not have norms, the results obtained cannot be compared with data from other schools. However, the industrial education teacher can study his success as a teacher by comparing results from semester to semester and from year to year.

10. Specific Uses of Tests.

The specific applications of tests to industrial education are discussed under the following general topics:

1. The measurement of class and pupil achievement.
2. The establishment of standards and norms of performance.
3. The motivation of learning.
4. The determination of efficiency of instruction.
5. The placement and guidance of pupils.
6. The evaluation of teaching materials and methods.

This broad scope of usefulness indicates that measurement is a fundamental factor in teaching industrial education subjects and that it is largely through the application of measurement to these subjects that adequate teaching methods and materials will be developed.

11. Measurement of Class and Pupil Achievement.

The very principles lying back of the construction of educational tests almost guarantee their usefulness to the classroom teacher in evaluating individual pupil and class accomplishment. The selection of the test items to cover the basic portions of the course of study which is or should be taught provides one basis for comparison. The form in which the test exercises are stated eliminates the personal equation of the individual teacher. The length of most of our approved standard tests guarantees consistency in the results of their use. The existence of norms and standards gives definite meaning to their scores. It is, therefore, a relatively simple matter for the classroom teacher to secure an accurate measure of the accomplishment of his class.

The use of a standard test in almost any selected subject makes possible the direct comparison of the individual pupils in the class on

an objective basis. The simple procedure of determining the average of the scores made by the class permits a direct comparison of this particular class with other, comparable classes in the building or system. Another very useful type of comparison is one which is frequently made between achievement at the beginning and at the end of a period of instruction. Each particular type of comparison serves its own purpose of assisting the teacher in determining the relative attainment and progress of his class.

12. Establishment of Goals of Attainment.

The fact that a test has been put through the process known as standardization gives it a distinct value in the classroom which an informal test does not have. The establishment of standards or norms for a test sets up in an objective way the goals to be attained in the course. The determination of whether or not a wood joint, a solder joint, a rope splice, a wire connection, a hem, a type of stitch, a drawing, a sample of freehand lettering is acceptable is not necessarily a matter of individual teacher judgment. In many of these fields objective standards on tests and scales establish these levels of attainment.

The comparison of results obtained from shop or laboratory projects with norms and standards for tests and scales gives the teacher an accurate indication of the achievement of his class in relation to other classes at the same experience or grade level. Experience shows that it is very helpful for an industrial education teacher to be able to evaluate his teaching success in terms of other teachers' accomplishments. If the results are consistently lower and the test items cover the course of study in a suitable manner it indicates the need for improved methods on the part of the teacher, or else reveals very low aptitude on the part of the pupils.

An example of test norms in the industrial education field is given in Table 8, which shows data on norms for the *Nash-Van Duzee Woodwork Test*. The norms (medians) show the average achievement scores of pupils on this test according to the number of minutes of instruction they have had. Any teacher capable of giving this test can compare the median of his class scores with the general average over the United States. In a large city the average accomplishment of different classes in the same school system may be compared in a similar way. Table 8 gives norms from the *Nash-Van Duzee Woodwork Test* I, Scale B.[1] This table shows the norms on the basis of semesters

[1] Nash, Harry B., and Van Duzee, Roy R., *Woodwork Tests*, The Bruce Publishing Company, Milwaukee, Wisconsin.

of work and amount of instruction in minutes. Thus the teacher will be able to compare the median achievement of his class with the achievement of other similar classes that have had the same amount of instructional time or that have been in the course the same number of semesters.

TABLE 8

SHOWING MEDIAN SCORE NORMS, BASED ON A NON-TIME AND A TIME SITUATION

Junior High School

	First Semester	Second Semester	Third Semester	Fourth Semester	Fifth Semester	Sixth Semester	Seventh Semester
End of semester work	1400 minutes	2400 minutes	3400 minutes	4600 minutes	6000 minutes	9000 minutes	17,000 minutes
Non-time median score	36	45	55	60	66	75	83
Time median score	44	47	53	58	64	71	80

Senior High School

Eighth Semester	Ninth Semester and up	Possible Score
25,000 minutes	32,000 minutes	
90	105	184
86	101	199

13. Motivation of Student Learning.

Tests and examinations have long been recognized by teachers as useful motivation devices. Many teachers have not realized, however, that the extent of this utility depends upon the character of the tests themselves. If the test is so constructed that it permits superficial thinking and shallow answers it stimulates precisely that type of work. If it calls for critical thinking, exact results, concise statements, careful evaluation of facts, then the force of the motivation is in the right direction. The use of even a moderately good test or examination may accomplish much in the way of stimulating proper habits of work on the part of the pupils. Sometimes even the mere administration of the test, or the knowledge on the part of the pupils that it is to be given, has a desirable effect. The greatest good, however, comes from the use of a carefully standardized test or scale, followed by the exact

location of individual pupil weaknesses and the application of corrective measures immediately after their discovery. The best experimental evidence shows that significant gains in pupil accomplishment accompany the sane use of properly constructed tests in such a way that the pupil himself is aware of his accomplishments and limitations.

14. Determination of Efficiency of Instruction.

These comparisons are interesting and often valuable as general guides, but if pupils are making low scores it is much more important to know where the scores are low. Is it in lettering, lack of textbook knowledge, poor technique, wrong type of instruction sheets, dull tools, or lack of interest? Just what are the conditions which cause the class to be lower than it should be on a standardized test? By a careful analysis of results it is often possible to determine weak points in the achievement of the class. A chart with the numbers of the test items in the standardized test on the left-hand side and the number of pupils getting the item correct on the right side is very useful for this purpose. Table 9 gives an example of this type of instructional analysis from a class of twenty eighth-grade boys as tested by Form B of the *Newkirk-Stoddard Home Mechanics Test.*[2] The test was given at the end of one semester of instruction. An examination of Table 9 shows that items 1, 6, 8, 12, 16, 20, 23, 24, 26, 27, and 34 are low in the numbers of pupils responding correctly. This analysis indicates that the jobs which correspond to these numbers probably were not taught effectively or at least were not properly mastered by the class. As a matter of fact, in this particular case, the main reasons for the ineffective teaching were lack of supplies for teaching the jobs properly, poor demonstrations, no supplementary references, and a lack of instructional time on the part of the teacher due to an unduly heavy teaching load in other branches.

15. Class Diagnosis.

Standardized tests of achievement are of value to industrial education teachers in determining the difficulties and abilities of the various members of the class. It is generally known that the background and abilities of the individual members of any class may vary widely. If the shop teacher is able to secure an accurate picture of the information and skills that the pupils already have when they enter the class, it will be of great value in placing the emphasis so the greatest instructional efficiency will result from the time allotted. This type of

[2] Newkirk, L. V., and Stoddard, George D., *Newkirk-Stoddard Home Mechanics Test,* Bureau of Educational Research and Service, Iowa City, Iowa, 1928.

TABLE 9

Analysis of Class Instructional Weakness in Home Mechanics

Newkirk-Stoddard Home Mechanics Test, Form B

Test Items	Number Correct Responses
1	6
2	18
3	20
4	18
5	15
6	3
7	17
8	2
9	20
10	18
11	15
12	6
13	15
14	14
15	18
16	5
17	3
18	20
19	18
20	2
21	18
22	17
23	1
24	3
25	18
26	4
27	1
28	14
29	18
30	17
31	15
32	14
33	16
34	2
35	13
36	11

information is especially useful to industrial education teachers who are teaching advanced classes, but it is valuable in all classes on the secondary level.

Even in small classes it is obvious that there is wide variability in the number of jobs that the pupils already know how to do and also

wide variability in the specific jobs. Table 10 shows this very clearly by data obtained by giving the *Newkirk-Stoddard Home Mechanics Test* to a class of nine eighth-grade boys at the University of Iowa High School to determine which items in home mechanics they already knew and to see where to put the instructional emphasis for each pupil.

TABLE 10

NUMBER OF ITEMS IN NEWKIRK-STODDARD HOME MECHANICS TEST EACH PUPIL ANSWERED CORRECTLY [3]

Pupil	Items	
	Form A	Form B
L	12; 15	1
T	1; 2; 4; 5; 14	2; 5; 14; 17; 32; 36
P	1; 2; 4; 5; 9; 10; 15; 30	1; 2; 4; 5; 10; 12; 13; 14; 16; 29; 34
S	1; 6; 4; 5; 22	1; 4; 15; 16; 28
D	2; 13; 33; 34; 35; 36	2; 4; 5; 15; 33; 34; 35
H	1; 2; 4; 5; 9; 12; 13; 23	1; 2; 6; 8; 10; 12; 14; 15; 33; 34
Z	1; 2; 4; 15; 36	4; 6; 12; 34
W	1; 2; 4; 5; 6; 8; 18; 30; 32; 34	2; 4; 13; 14; 16; 17; 26; 28; 29; 33
M	2; 6; 7; 8; 9; 12; 17; 26; 29	2; 3; 4; 6; 8; 9; 10; 11; 13; 16; 18; 19; 21; 23; 28; 29; 33; 34; 36

Out of the seventy-two items in the two forms of the test, only twenty-three were not answered correctly by some of the pupils. The highest score that any one received was twenty-eight jobs right (pupil M). The results are very valuable from the standpoint of instructional efficiency because this pupil will not have to spend time repeating material with which he is familiar. In the case of pupil L, who scored only three right, the test has identified an individual who needs careful attention. To the pupil himself it is a clear indication of the need for more instruction.

Information of this type is valuable not only for increasing instructional efficiency, but also for motivating the pupils on their proper level of accomplishment. Unless the teacher has previous knowledge that pupil M knows how to do twenty-eight of the jobs specified he is quite likely to waste his own and the pupil's time through useless repetition. This usually results in building up bad habits of work on the

[3] Newkirk, L. V., *Validating and Testing Home Mechanics*, University of Iowa Study in Education, University of Iowa, Iowa City, Iowa, Series 201, 1931, pp. 30-31.

part of the pupil. Furthermore, the teacher might assume that pupil M did not know how to do any of the tasks in the test when as a matter of fact he knows much of what he has to learn. For example, pupil L who knows three jobs might learn twenty more and his score would then be twenty-three, and pupil M who knows twenty-eight might learn ten more and have a score of thirty-eight. The boy who had learned twenty would have accomplished more, but the final score would not indicate that he had accomplished twice as much. In fact, it would give the impression that he had accomplished fifteen less. This merely illustrates the need for giving industrial education tests at the beginning of a course to discover what is already known, during the semester for indications of progress and for motivation, and at the close of the semester to measure accomplishment and growth.

16. Individual Pupil Diagnosis.

Closely related to the measurement of class and individual levels of accomplishment is the diagnosis of individual learning difficulties of certain pupils in the class. Just as in the other instructional fields, the teacher may assume that these pupils are naturally slow or do not try. It frequently occurs, however, that upon closer examination these slower pupils have many learning difficulties which can be corrected by the application of proper remedial teaching. The possible causes of these difficulties are numerous; they may be one or more of the following: malnutrition, defective eyesight, difficulty in hearing, poor reading ability, poor technique in manipulation of some or all tools, inability to adjust tools, inability to read a working drawing, ignorance of sizes of tools, unfamiliarity with related mathematics, low mechanical ability, emotional maladjustment, social maladjustment, and low intelligence. These difficulties are usually obvious in extreme cases, but the majority of the pupils in the class may have one or more of the difficulties which will seriously affect his ability to profit from the instruction. It is on this account that the industrial education teacher needs as much professional information about his pupils as it is possible to secure in order better to adapt his instruction to the individual differences and abilities of his pupils.

The efficient shop teacher must know how to test many factors other than those that relate directly to achievement in industrial education. Fundamentally he is a teacher of individuals and not a teacher of drawing, woodwork, metal work, electricity, printing, or auto mechanics. Table 11 illustrates types of information which industrial education teachers will find useful in their teaching and guidance activities.

TABLE 11

DESIRABLE TYPES OF PROFESSIONAL INFORMATION

Name......John J....... Grade......10......

Test	Score
Intelligence test	112
Reading	40
Language	20
Spelling	55
Writing (quality)	60
Mathematics	75
Mechanical aptitude	120

Hypothetical grade norms for the test scores in Table 11 are given as follows:

Grade	Intelligence Test	Reading	Language	Spelling	Writing	Math-ematics	Mechanical Ability
7	90	40	20	40	60	30	70
8	95	45	32	45	65	40	80
9	100	55	45	50	72	50	100
10	110	60	50	71	80	75	120
11	122	65	60	75	86	80	135
12	136	75	65	82	90	84	150

The information in Table 11 gives the achievements obtained by a tenth-grade pupil on a number of tests, and the hypothetical grade norms indicate what the pupil's level of achievement should be. The levels of accomplishment are indicated graphically in Fig. 2.

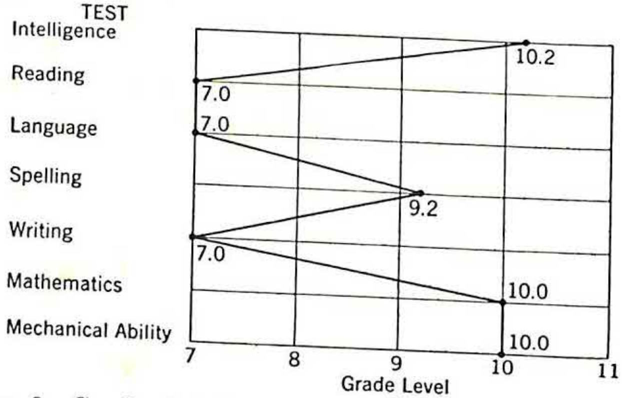


FIG. 2.—Graph of Achievement According to Grade Level.

The profile chart shows that John J. has intelligence slightly better than a tenth-grade pupil, the reading and language ability of a seventh-grade pupil, spelling ability a little above that of a ninth-grade pupil, writing quality equal to that of a seventh-grade pupil, mathematical and mechanical ability equal to that of a tenth-grade pupil. Assume further that this pupil is in the tenth grade in electrical shop and is making slow progress. The teacher in the electric shop is using instruction sheets and related reference materials as supplementary teaching devices. The students are required to do considerable reading and to write out the answers to the questions on the individual instruction sheets. By studying the chart in Fig. 2 it is obvious why this pupil has difficulty in making satisfactory progress. He cannot read well and is a poor writer, although he has intelligence and mathematical and mechanical ability adequate for doing good work in the course. The remedy here is special instruction in reading and language with additional emphasis on writing legibly.

All the illustrations used here have been on the basis of grade norms because they are easy to compute and illustrate the different levels of accomplishment. However, many industrial education teachers may wish to classify pupils on the basis of ability to learn, and reveal progress by using mental ages and achievement ages of the pupils in their classes. The various types of norms are discussed in Chapter XV.

17. Gradation and Guidance.

Tests of intelligence and mechanical and special aptitudes are of value to supervisors and teachers of industrial education in classifying pupils with approximately equal learning power. In a large school system where there are several sections of a class, it is usually considered desirable for instructional purposes to classify pupils into groups of about equal learning abilities. It is easier to meet the individual learning difficulties of a group of pupils if they have about the same general intelligence. In the small school it may not be possible to divide the pupils into instructional groups of approximately equal learning ability, but usually the classes are small and the teacher has more time for individual instruction. The industrial education teacher may divide classes on the basis of either mechanical ability or intelligence. Both these factors are important in shop instruction, but they do not correlate highly. Some pupils have low mechanical ability and high intelligence, others high mechanical ability and low intelligence, as indicated by intelligence tests. Scores on intelligence tests and scores on mechanical-ability tests usually have a positive correlation

of between .20 and .30. Of course, the large majority of pupils have about average mechanical ability and average intelligence. If an industrial education class is selected on the basis of scores on intelligence tests the result will be a class with similar intelligence ratings but with variable mechanical aptitude. If selected on the basis of mechanical aptitude the intelligence ratings will be variable.

In the first two years of the junior high school where information about tools, materials, and industries is considered more important than acquiring outstanding tool skill, it seems desirable to section classes on the basis of intelligence scores, because of the nature of the learning problems. In advanced courses where trade training and the acquiring of trade skill are the dominant objectives, it probably is better to classify pupils on the basis of mechanical aptitude, since that is of vital importance in acquiring outstanding skill in manipulating tools and materials. In either case, it would be desirable to have both ratings for use in adapting instruction to the individual difficulties of the pupils.

Tests of mechanical aptitude or mental ability are usually administered by the supervisory officers in the school or by persons especially trained in administering tests. If this practice is followed, the industrial education teacher can frequently get the necessary information from the central office of the school. However, the needed information is not always available, and oftentimes it is necessary to check uncertain scores or to test pupils who have recently entered school or for whom test scores are not available. Industrial education teachers also need information about giving and scoring tests of special abilities so that the scores and the implications for their use will be clear.

The chief danger in using test scores for gradation or guidance purposes is that they may not be interpreted in the light of their true meaning. The scores from the best educational measures are not so reliable for individual diagnosis as they are for indicating general trends or levels of accomplishment. It has been found that test scores which are very high or very low are most likely to be in error. The combined scores of several similar tests can be used with more certainty in diagnosis of pupil difficulties than any one score. Before very high or very low scores are used they should be rechecked by giving similar tests or other forms of the same test. Scores obtained on carefully prepared educational tests are more accurate than the teacher's subjective judgment, but they are not accurate enough to be considered final and used dogmatically.

Teachers of industrial education should be very careful not to confuse the purposes of aptitude and special-ability tests with achieve-

ment tests. A pupil who has a high score on tests of intelligence and mechanical ability, other factors being equal, should do good work in industrial education courses. The fact that a pupil has these abilities does not necessarily mean that he should receive a high mark. Regardless of a pupil's abilities, he should be marked on the basis of actual achievement in the course taken. Standardized and teacher-made tests should be utilized for measuring achievement and the results used as a major factor in assigning shop marks. Tests of special abilities are valuable in guidance, in classification of pupils, and in pointing out individual pupil difficulties. They are not of particular value in measuring the amount of information or skill acquired in industrial education courses.

18. Tests in Research.

One of the obligations of a teacher to his profession is to discover new truths which can be applied for the improvement of work in his chosen field of endeavor. Carefully constructed educational tests can be used to discover new and better ways of organizing and teaching industrial education. It does not seem likely that a scientific method of instruction can be developed in any instructional field without suitable measures of achievement and abilities. The following are examples of a few of the problems which could be solved in part through the use of adequate tests.

1. What are the relative values of different teaching methods for industrial education subjects (use of demonstrations, instruction sheets, class instruction, individual instruction)?
2. What type of shop organization is most effective (composite, unit)?
3. How much instructional time should be given to lecture, demonstration, and individual instruction?
4. What types of individual instruction sheets are most effective at different grade levels?
5. What is the proper size of a class in drawing, sheet metal, machine shop, foundry, woodwork, auto mechanics, printing, and the general shop?
6. What is the most economical length of period to be used in industrial education instruction?
7. What is the most effective classification of instructional materials in industrial education courses on the basis of grade accomplishment?

SUMMARY

Educational measurements have the following general uses in industrial education: to measure class and pupil achievement, to establish standards of performance, to motivate learning, to diagnose pupil learning difficulties, to mark and promote pupils, to classify pupils according to abilities, and to study the effectiveness of teaching methods.

Educational measurement is a fundamental factor in teaching industrial education. Standardized and teacher-made tests are valuable in measuring achievement. Aptitude tests are valuable in guidance and diagnosing individual difficulties. Teachers need a great deal of professional information about their pupils other than measures of achievement if their courses of instruction are to be effectively adapted to individual needs of their pupils.

SUMMARY EXERCISES FOR DISCUSSION

1. List the major factors which would make it difficult for the classroom teacher to construct tests which will have the merits of carefully constructed standardized tests.
2. Enumerate and illustrate the six main uses of tests in industrial education.
3. Show how tests of intelligence and special aptitudes may be used for gradation and guidance purposes.
4. What is the teacher's responsibility for the use and interpretation of standard tests and scales in the classroom and shop?

SELECTED REFERENCES

GREENE, H. A., and JORGENSEN, A. N., *The Use and Interpretation of Elementary School Tests.* New York: Longmans, Green and Company, 1935.

BREWER, JOHN M., *Cases in the Administration of Guidance.* New York: McGraw-Hill Book Company, 1929.

MONROE, W. S., *The Theory of Educational Measurements.* Boston: Houghton Mifflin Company, 1923.

NEWKIRK, L. V., and STODDARD, GEORGE D., *The General Shop.* Peoria, Illinois: The Manual Arts Press, 1929.

ODELL, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

RUCH, G. M., and STODDARD, GEORGE D., *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927.

SMITH, H. L., and WRIGHT, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928

SYMONDS, P. M., *Measurement in Secondary Education.* New York: The Macmillan Company, 1927.

WILSON, G. M., and HOKE, K. J., *How to Measure* (Revised). New York: The Macmillan Company, 1928.

CHAPTER IV

SELECTION AND EVALUATION OF TESTS

I. CRITERIA FOR INDUSTRIAL EDUCATION TESTS

Several characteristics of a good test should be considered by the shop teacher in evaluating published tests or tests of his own construction. The most important of these are validity, reliability, objectivity, adequate norms, the existence of duplicate and equivalent forms, ease of administration, and economy. An understanding of these factors will do much to insure the selection or construction of a test suitable for the testing problem at hand.

19. Validity.

The general concept of validity in a test may be made clear by thinking of the conditions set up by the test as a small sampling of a larger life situation. At the outset it is assumed that the field which the test samples is of some real importance. If this is the case, then the more nearly the conditions set up in the test itself duplicate the larger situation as found in life the more valid it becomes. For example, it would doubtless be possible to prepare a laboratory test designed to measure one's ability to handle an automobile in heavy city traffic and one's reactions to the situations encountered there. It would be much more practical (valid) to bring the subject into direct contact with a bit of heavy traffic and determine exactly how he does react to it.

From the point of view of the classroom teacher, validity usually is concerned with the question of whether the materials tested are actually of real significance, and whether the pupil has had any adequate opportunity to master the facts tested as a result of his contact with the course of study taught. *Validity may be defined as some type of objective expression of the degree to which the particular measuring instrument measures what it is supposed to measure.* That is to say, a test which is designed to measure ability to read blueprints after a short period of training, and later is found to be a better test of general intelligence, would be considered to be lacking in validity for the purpose for which it was designed. Validity is usually expressed in terms of the correspondence of results obtained from the

particular measuring device under consideration and other, similar instruments of previously determined validity. Very often it is impossible to secure measures from other instruments of known validity. In these cases it is a common practice to refer to estimates or judgments of individuals who have had an opportunity to evaluate in a rather definite way the abilities of the individuals involved in the validation study. Frequently validity is determined by the extent to which a test calls into play the skills and abilities which experienced observers consider fundamental to success in the given field. The validity of many of the items in the *Newkirk-Stoddard Home Mechanics Test* is dependent to a large degree upon the agreement of certain teachers, supervisors, and other qualified authorities that the processes called for are the significant ones.

The validation of the content of this test was achieved in part by the pooled judgment of experienced teachers, home owners, and tradesmen. The home owners indicated the projects and content which they believed to be important in the maintenance and operation of the home. The teachers of home mechanics in 75 schools marked the jobs which they considered most important. In developing the procedure type of question used in the test, it was necessary to have the procedures checked against good trade practice by tradesmen. Table 12 gives the ten most frequently occurring home mechanics

TABLE 12

Ten High-ranking Home Mechanics Jobs According to 100 Home Owners

Job	Frequency
1. To sharpen knives	98
2. To install a pair of hinges	98
3. To put new screen on a window or door	95
4. To connect batteries	95
5. To shape the point of a screw-driver	95
6. To wash a window	95
7. To use glue for general repair	94
8. To regulate a watch or clock	94
9. To fire a furnace	94
10. To locate a blown fuse and replace	94

projects according to the judgment of 100 home owners living in small towns in the middle west. Table 13 gives the ten highest-ranking jobs in home mechanics according to the judgment of 75 teachers of the subject. Table 14 gives a procedure rearrangement question taken from the *Newkirk-Stoddard Home Mechanics Test,* the numbers

in the parentheses indicating the best trade procedure according to the five tradesmen who judged it.

In many of the achievement tests, validity depends to a large degree upon the opportunity which the pupil has had to master the information covered by the test. The validity of a test may be thought of as being *general,* or it may be considered as being *specific.*

TABLE 13

TEN HIGH-RANKING HOME MECHANICS JOBS ACCORDING TO 75 HOME MECHANICS TEACHERS

	Job	Frequency
1.	To make suitable splices, tops, and terminals in electric wires	64
2.	To tin a soldering copper	63
3.	To wire an electric-light socket	63
4.	To mend leaks in kitchen utensils	62
5.	To make an extension cord	62
6.	To wire simple bell circuits	61
7.	To apply stain and filler	61
8.	To apply varnish	60
9.	To cut glass to size	60
10.	To repair leaking compression faucet	57

TABLE 14

A REARRANGEMENT QUESTION WITH THE ANSWER AS APPROVED BY 5 TRADESMEN

To Cut a Piece of Pipe:

Procedure:
(1) Set the cutter on the mark.
(2) Ream the end.
(3) Determine the length of pipe.
(4) Cut the pipe.
(5) Measure and mark.
(6) Adjust the pipe cutter.

(3) (5) (6) (1) (4) (2)

Many tests are undoubtedly valid in a general sense but are lacking in validity in a specific sense. For instance, a survey test designed to secure a general bird's-eye-view of achievement in a particular subject must be validated in terms of its ability to test for the basic items found in the courses of study, not of a single class in which certain points of view and certain facts have been emphasized, but of the many different schools in which it may be used. For his own particular class a teacher may easily construct a test which will have much greater specific validity for his purposes and his point of view

than any type of commercial standardized test could possibly have. Recently considerable recognition has been given to this phase of validity in tests by providing for the classroom teacher source books of objective test exercises in a number of subject matter fields.[1] Such material permits the classroom teacher to secure a relatively high specific validity for his tests and quizzes.

20. Reliability.

The reliability of a test may be thought of as the consistency with which it performs. In a certain sense this matter of consistency of performance of a test arises from two factors, the adequacy of the sampling represented by the test, and variations in the human response itself which have nothing to do with the content of the test. The first of these can be controlled somewhat by selecting the test items carefully and extensively from the field which it is supposed to measure. The principle of sampling may be illustrated by the practice of the large producers of ore. Obviously it would be impossible to examine and test every cubic foot of the ore in every carload. It is a simple matter, however, for specimens of the ore to be taken from different parts of the car and from different cars. These specimens are carefully mixed together and subjected to the tests which determine the quality and price of the ore. This process is called *sampling.* If only one specimen were taken from each car there would always be the possibility that the ore at that particular spot might have been unusually rich or poor. Taking more and more specimens increases the likelihood that the resulting sample will be truly representative of the ore in the car. In a similar way, increasing the number of samples taken in a testing situation makes it more likely that some important phase of the subject may not have been missed or given the wrong emphasis, or that some interfering human factor may have been operating at the time the samples were taken.

The accompanying diagram illustrates the effect of sampling on the reliability of a test over a limited field of information. Each of the small rectangular spaces in Fig. 3 represents an item which a student has had an opportunity to learn. The thirty shaded portions represent the items which he has actually learned. The ten unshaded spaces are items which he has not learned. In this illustration the student has a mastery of 75 per cent of the items. Now let it be assumed that a test over this material is prepared comprising ten items, numbers 1, 2, 3, 6, 7, 8, 10, 12, 16, and 22. If these items are

[1] Kirkpatrick, J. E., and Greene, H. A., *Pupil-Teacher Handbooks of Objective Exercises in High School Physics.* Bloomington, Illinois: Public School Publishing Company, 1930.

selected, this individual will fail on six of the ten, making a percentage score of 40 per cent. However, if ten other items, as numbers 5, 9, 13, 17, 20, 24, 27, 31, 34, 37 are selected over the entire range of his field of information he should be able to answer nine of the ten correctly. From this it is clear that it makes a distinct difference where the sampling is taken. It may cover material learned but may come from too limited a portion of the total field to be truly representative of the individual's accomplishment. In this illustration, if the even numbered items are chosen, the pupil would probably fail on nine of the twenty items. If the odd-numbered items are chosen he should fail on only one. This results in a variation in his score from 45 to 95 per cent. As the number of items chosen for inclusion in the test is increased the pupil's scores on the test exercises more nearly approach the actual amount of his information in the field. It thus becomes apparent that the extent of the sampling is also an important feature of reliability in a test.

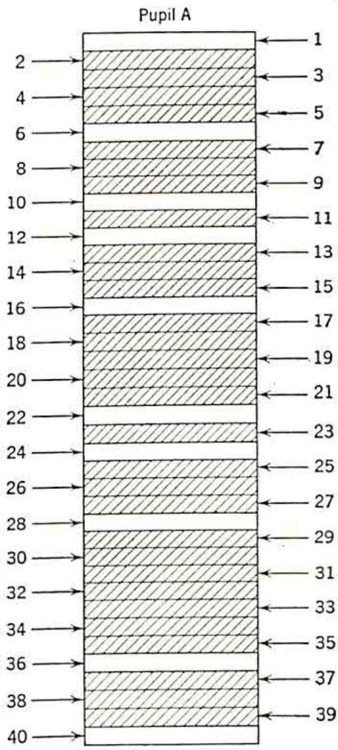


FIG. 3.—The Principle of Sampling

As the reliability of a test is increased, either through more extensive or more representive sampling, the operation of chance variations such as temporary disturbances, breaking a pencil, and the like, is minimized. Similarly, increasing the sampling of the test by increasing the number of different times the pupil is required to respond to it tends to limit the effect of physical disturbances, fatigue, emotional stress, etc. Thus, practically every attempt to increase the consistency with which educational tests measure abilities and achievements results in the increase in the length of the test and testing period. It is becoming increasingly clear that complex fields cannot be measured reliably by means of brief tests.

The reliability of a published test should be given in the manual of directions accompanying the test. This information is given as a coefficient of reliability. The coefficient of reliability is a statistical expression of the consistency of performance of the test and how much reliance may be placed on scores obtained from its use. Reliability

coefficients are indicated in decimal fractions of 1, usually ranging in value from .40 for tests of low reliability to .95 for tests of very high reliability. It is improbable that a test will be entirely inconsistent or perfectly consistent. However, it is difficult to state exactly how low or how high a coefficient of reliability of a test should be for the test to be of value. Much depends on the type of test and how it is to be used. The data in Table 15 will give some appreciation of a suitable coefficient of reliability.

TABLE 15

RELIABILITY COEFFICIENTS

Coefficient of Reliability	Rating
.50- .60	Very low
.60- .70	Low
.70- .80	Fair
.80- .85	Average
.85- .90	Good
.90- .95	High
.95-1.00	Excellent

In general, it is doubtful if a test should be given very serious consideration if it has a reliability coefficient of less than .80 when stepped up by the application of the Spearman-Brown technique from correlations on chance-half samplings of the exercises. Some experience with long and quite reliable tests in the field of silent reading indicates that reliability coefficients which range as high as .95 when computed on the odd-even basis drop as low as .90 when results from the two equivalent forms of the test are correlated. Reliabilities of .90 based on the actual relationships between two equivalent forms of the test may be considered very high. Additional discussion of the meaning and significance of reliability, as well as a more complete explanation of the methods of securing and computing reliability data, are given in Chapter XIV.

21. Objectivity.

Objectivity is an important quality in a test exercise since it contributes indirectly to validity and reliability. *Objectivity is that quality in a test exercise which makes for the elimination of the personal judgment of the person who scores it.* Since this means greater accuracy in the grading of such items it naturally indicates greater reliability in the measurement of whatever qualities are being meas-

ured. Objectivity is a function of the form in which the test items are stated. In general, objective-test items are so formulated that only one correct response satisfies the conditions of the exercise. Recall, true-false, multiple-answer exercises are all illustrations of objective forms. These and other common forms of objective exercises are described and illustrated on pages 107-123 of this book.

22. Ease of Administration.

The speed, accuracy, and simplicity with which an educational test can be given in the classroom, though not a major criterion for tests, is nevertheless one that is worthy of some practical consideration. There is a very definite tendency in modern test development to recognize the teacher's and the supervisor's problems by making the tests easy to administer and simple to interpret.

A significant phase of the administrability of a test is the examiner's manual which accompanies it. The manual should contain a clear statement of the qualities measured by the test. It should provide concise and simple directions for giving the tests so that they may be followed verbatim by the classroom teacher. It should provide the critical user of the test with an adequate explanation of the methods by which the validity and the reliability data were obtained, as well as a concise statement of the meaning of these data in relation to the tests themselves. A convenient form of answer key or simple stencil for scoring should be supplemented by brief explanatory directions concerning the methods of scoring the tests. Simple illustrations of the methods of interpreting the results should be given in the manual; and if the field is one in which a follow-up program of remedial or corrective instruction is possible, brief suggestions for such work should be made.

The better standardized tests provide the user with carefully formulated statements on the following types of items, all of which tend to protect the pupils and the teacher against the faulty administration of the tests and wrong and uncritical interpretations of the results:

1. Number of parts in the test.
2. Directions for each part or division of the test.
3. Sample or fore exercises to acquaint the pupils with the method of work.
4. Directions for stopping at the end of a test part or for turning the page when necessary.
5. Definite statements of time limits where required.
6. Directions for scoring the tests.
7. Explanations of the norms or standards of accomplishment.

8. Statement of the total possible scores on each test part and the method of securing them.
9. Explanation and illustration of method of interpretation of results in terms of apparent instructional needs.
10. Suggestions for definite remedial attack on the weaknesses revealed by the tests.

23. Norms and Standards.

Norms and standards, although frequently used as synonymous terms, are not exactly identical in meaning. Norms represent actual levels of accomplishment for specified groups of individuals. Standards are usually considered as representing goals to be attained. One is what children or pupils are actually able to do; the other is what the teacher should strive to have them do. Practically all present-day tests are supplied with norms rather than standards.

Norms are usually based upon the average or median accomplishment of large numbers of pupils grouped by ages or by grades. Grade norms result from the classification of the pupils by grades. Age norms result from the classification by ages. The norms, therefore, furnish the teacher with a definite basis for anticipating what given groups of pupils may be expected to achieve under ordinary schoolroom conditions. They thus afford the basis for the practical interpretation and evaluation of the testing program and of the classroom instruction under analysis.

In general, norms, in order to have sufficient reliability for practical classroom use, must be based upon rather large and carefully sampled populations. There is a growing tendency, however, to base standard test norms upon smaller groups of specially selected cases. In the past it has been assumed that the inclusion of a large population of unselected eighth-grade pupils would afford the best basis for an eighth-grade norm in a specific test field. Evidence brought forward by Crawford [2] indicates that this is not necessarily true. If there were a perfect balance in the proportion of pupils in a given grade who are retarded and accelerated as to mental ability and school progress, such an unselected group might provide suitable and representative norms. The actual evidence shows, however, that in the typical school grade the retardation (retarded progress) actually exceeds the acceleration by a ratio of four to one and sometimes six to one. Accordingly, the norms based on unselected cases are not typical of the actual achievement of the normal individual in the

[2] Crawford, J. R., *Age and Progress Factors in Test Norms*, University of Iowa Studies in Education, Vol. 9, No. 4, June, 1934. University of Iowa, Iowa City.

group. Serious consideration is being given by modern test workers to the more exacting control of these variables. A number of the newer tests are reporting norms based upon smaller groups of individuals selected for their normality for the group of which they are a part. In a number of cases age norms are based upon the results of grouping the individuals by means of mental-test scores. This procedure alone takes care of the serious difficulties arising when the grouping is by chronological age. To the extent that mental tests are standardized accurately, a twelve-year-old (mental age) individual is a twelve-year-old wherever he may be found. Every teacher knows, however, that a chronological twelve-year-old in the fifth grade is quite unlike one in the seventh grade.

24. Mechanical Features.

The mechanical features of a test frequently operate definitely to affect its ease of administration in the classroom. They are largely the result of the editing and printing of the test. The paper should be of good quality, preferably white bond. The illustrations should be clear-cut and easily identified with the content they are supposed to illustrate. The page size, the length of line, and the size of type used are also mechanical features which may influence the usefulness of a test.

25. Economy.

Other things being equal, economy as a criterion for standard tests should undoubtedly be listed last. In the final analysis, any test which takes up class time may be counted expensive. Moreover, the cheap test is not always the most economical; in fact, quite the reverse is apt to be true. Tests costing at the rate of 50 cents per hundred and yielding results of limited validity and low reliability might readily be much more expensive than tests costing six times as much but having validity indices of .80 and reliability coefficients of .93. It is not at all unlikely that in the near future educational tests will be evaluated in terms of the number of units of valid and reliable information yielded per unit of cost.

A modern tendency in test development involves the introduction of certain economy features in the booklet, either through the use of automatic scoring devices or through design which permits the repeated use of the test booklets. Such mechanical features in a test are quite acceptable so long as they do not force the test into too badly crowded a page, or do not interfere with the validity and reliability of the measurement which would otherwise be attained.

26. Number and Equivalence of Forms.

The more useful educational tests are those which exist in multiple forms. The forms of a test are secured by preparing two or more arrangements of similar but not identical test exercises and assigning them to different test booklets. These multiple forms should normally be approximately equal in difficulty, not only in terms of the total scores earned by groups of equally able individuals, but also in terms of the ratings of items in the different levels of the test.

II. EVALUATION OF TESTS

27. Test-Rating Scales.

In the foregoing discussions no attempt has been made to evaluate the items mentioned in the criteria but merely to explain and discuss the types of information which are of value in selecting a published test or in judging the value of an objective shop test. However, a number of rating scales have been developed in which the different factors are weighted and give a test a rating on the basis of 100 points. These scales are useful to all teachers in selecting tests but are especially valuable to the inexperienced teacher until he becomes accustomed to judging the different items. The Otis test-rating scale is reproduced here.

Otis Test-Rating Scale [3]

Manual (5)
Validity (15)
Reliability (10)
Reputation (5)
Ease of administration (total 15)
- (a) Preparation (4)
- (b) Time limits (4)
- (c) Explanation needed (3)
- (d) Alternative forms (4)

Ease of scoring (total 15)
- (a) Objectivity (10)
- (b) Time required (3)
- (c) Simplicity (2)

Ease of interpretation (total 15)
- (a) Norms (5)
- (b) Directions for interpreting (4)
- (c) Class record (1)
- (d) Application of results (5)

Convenient packages (5)
Typography and make-up (5)
Test service (10)

Total 100

[3] Otis, A. S., "Scale for Rating Tests," *Test Service Bulletin* No. 13. Yonkers: World Book Company. 6 pp.

III. SUMMARY

The teacher should select his tests with care in order to get full value for the money and time expended for testing purposes. The chief factors to be considered in selecting a test are validity, reliability, objectivity, norms, multiplicity and equivalence of forms, ease of administration, and cost. Validity is the degree to which a measuring instrument measures the thing it purports to measure. Reliability is the consistency of performance of the test itself. Objectivity is determined by the form of exercise, which in turn controls the number of acceptable answers for each question. Norms are the median or average performance of pupils of different ages or grades as determined by the testing of large numbers of individuals. Standards are desirable ultimate goals of attainment. Tests should be economical of time and money. The Otis test-rating scale is suggested as a useful guide for the shop teacher in the selection of commercial tests.

SUMMARY EXERCISES FOR DISCUSSION

1. Formulate a concise definition of each of the major criteria for the selection of an educational test.
2. Illustrate at least three types of procedure which may be used in the validation of industrial education test items.
3. Show by means of a concrete illustration how sampling affects the reliability of a test.
4. What are some of the most effective devices for making test exercises objective?
5. In your judgment, is the apparent difference between norms and standards a matter of any practical significance?
6. Secure at least one complete sample set of standardized tests suitable for use in industrial education classes. Examine the tests, the manuals, keys, and other supplementary material and rate the test in detail, using the *Otis Score Card for Rating Tests.* Consider each item in the light of the explanations given with the score card, and assign a value to each point in proportion to its apparent merit.

SELECTED REFERENCES

Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests.* New York: Longmans, Green and Company, 1935.

Odell, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

Odell, C. W., *Traditional Examinations and New-Type Tests.* New York: The Century Company, 1928.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929.

Ruch, G. M., and Stoddard, Geo. D., *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

Wilson, G. M., and Hoke, K. J., *How To Measure* (Revised). New York: The Macmillan Company, 1928.

CHAPTER V

MEASURABLE FACTORS IN INDUSTRIAL EDUCATION

28. Factors Related to Mastery in Industrial Education.

Industrial education teachers, in common with teachers in most other subjects, need a great deal of precise information about the pupils in their classes if they are to work effectively. Most of the data needed can be obtained through measures which are much superior to the teacher's unaided judgment. Some of these measuring instruments are still to be developed, and it is certain that many now available need refinement. It is the purpose of this chapter to point out some of the measurable factors in industrial education. These factors present a real challenge to teachers of industrial education. Educational guidance and shop instruction both doubtless would be markedly improved through the measurement of these factors and the wise use of the results in the classroom. These measurable factors are enumerated in Table 16, and each will be discussed in this chapter.

TABLE 16

MEASURABLE FACTORS IN INDUSTRIAL EDUCATION

Measurable Factors	Comments
1. Information	Factual information about tools, materials, and vocations. (Oak is a cabinet wood; the micrometer is an instrument used to measure in thousandths of an inch; outside paint contains oil.)
2. Quality	Evaluation of the product of manipulative work in the light of tool, instrument, or machine operations (drawing, hammering, house wiring, bookcase, cement lawn pedestal).
3. Technique	Evaluation of skill in manipulating tools, instruments, or machines in executing tool operations (method of using a plane, a compass, a lathe).
4. Speed (Rate of response)	The time required to accomplish a piece of work employing commercial standards (time required to make a drawing, a table, wire a house).

TABLE 16—(*Continued*)

MEASURABLE FACTORS IN INDUSTRIAL EDUCATION

Measurable Factors	Comments
5. Reading technical symbols	Ability to read working drawings, wiring diagrams, architectural drawings, etc.
6. Reading	Ability to read and comprehend instructions or related information from the printed page.
7. Spelling	Evaluation of ability to spell common words and necessary technical terms.
8. Mathematics	Evaluation of mathematics required in the various shop courses (woodwork, drawing, sheet metal, machine shop, home mechanics).
9. Appreciation of industrial products	Evaluation of ability to rank industrial products according to merit (furniture, electrical devices, finishes, automobiles, houses, radio).
10. Planning	Evaluation of ability to develop a suitable plan for doing a job (building a lawn bench, a dog house, a fence, a radio, etc.).
11. Language	Ability to use correct English in written and oral form.
12. Inventiveness	Ability to see new relations and develop devices and machines for the improvement of society.
13. Personality Traits	Rating of traits generally recognized as essential to success (industry, cooperation, consideration for others, self-reliance, aggressiveness).
14. Mechanical aptitude	Natural aptitude for manipulating mechanical devices and an understanding of their operation.
15. Intelligence	Ability of an individual to learn as measured in terms of the extent a pupil has acquired a number of specific and largely unrelated abilities.

These fifteen factors are obviously closely related to mastery in industrial education. To the extent that they are basic they should represent the framework of measurement in this field. It is probable that they can be measured most effectively and the results used to

best advantage when each is tested separately. It should not be unduly difficult to construct devices for the measurement of information, tool techniques, quality, and speed in more or less isolated forms as a basis for a real analysis of achievement in this subject. It is believed, therefore, that these categories are of sufficient importance to warrant the elaboration and discussion of each in order of appearance in Table 16.

Information. Industrial education courses contain a great deal of knowledge about tools, machines, materials, and industries. This is especially true in the junior high school where a great deal of emphasis is placed on outlook and cultural training and less on the strictly vocational aspects of the subject. The vocational courses also contain much instructional material which is designed to develop knowledge about tools and materials rather than skill in the actual modification of materials through the use of tools and machines. This information or knowledge is similar in all its psychological aspects to knowledge in other subjects of the school curriculum and can be measured effectively with the common type of objective examination. In fact, the majority of the tests of achievement which to date have appeared in industrial education are measures of information. The same techniques which have been used for constructing tests in other subjects can be applied with only slight modification. However, there is still a great deal of work to be done before an adequate supply of measures of information will be available for the content of the varied subjects of industrial education. Some indication of the extent of this problem will be gained from an examination of the accompanying outline of informational items on woodworking adapted from the American Vocational Association committee's report[1] on standards of accomplishment in the industrial arts.

EXAMPLES OF THE INFORMATIONAL CONTENT OF INDUSTRIAL EDUCATION

Woodworking

A. Lumber.
   1. Know the principal characteristics, the working qualities, the principal uses, and the sources of supply of the following woods: pine, cypress, oak, walnut, birch, maple, mahogany, red cedar, hickory, gum, chestnut, poplar.
   2. Know the methods of cutting and milling lumber.
   3. Know how lumber is dried.

[1] Report of American Vocational Association Committee on Standards of Attainment in Industrial Arts, *Bulletin of American Vocational Association, Industrial Arts Section,* pp. 29-31, December, 1931.

4. Know the effect of moisture on wood.
5. Know the standard dimensions of lumber and how classified.
6. Know the nominal and the actual dimensions of lumber.
7. Know how veneer and plywood are made, and their uses.

B. Finishes.

1. The object of finishes.
2. The kinds of finishes in common use; such as stain, oil, wax, shellac, varnish, lacquer, enamel, paint.
3. The durability of different finishes.
4. The conditions or places in which various kinds of finishes may be used to advantage.
5. Materials from which finishes are made.

C. Glue.

1. The kinds of glue.
2. The preparation of glue.
3. The conditions and requirements in use.

D. Nails, brads, and fasteners.

1. The kinds of nails.
2. The uses of the different kinds.
3. The size of nails.
4. How nails are sold.
5. How nails are manufactured.
6. Sizes of brads and how sold.
7. Size, kinds, and uses of corrugated fasteners.
8. Sizes and uses of clamp-nails.

E. Screws.

1. The kinds of screws.
2. The uses of the different kinds.
3. How the sizes and kinds of screws are indicated.
4. How they are sold.

F. Sandpaper and steel-wool.

1. The kinds of sandpaper.
2. Grades of sandpaper.
3. Principal uses.
4. Grades and uses of steel-wool.

G. Design of furniture.

1. Is it adapted to the use for which it is indicated?
2. Is it structurally good?
3. Is it well made?
4. Are the structural members in good proportion?
5. Does it have an appearance of stability?
6. Is the structure as a whole well-proportioned?
7. Are the outlines pleasing?
8. Is it well finished with an appropriate finish?

H. Manufacture of wood products.
   1. The location of important manufacturing concerns.
   2. The division of labor in industry.
   3. The use of automatic machinery.

I. Joints.
   1. Types of joints, where used, and why.

J. Hardware.
   1. Types of hinges and their uses.
   2. Types of latches and where used.
   3. Types of locks and where used.
   4. Types of nails and where used.
   5. Special types of fittings.

K. Abrasives.
   1. Kinds of grinding and sharpening stones, their grades and uses.

The American Vocational Association committee on standards of accomplishment in industrial arts has aptly referred to knowledge of the informational type as "the things you should know." The example adapted from the committee's report are not given as necessarily complete either by the committee or the present authors, but they are suggestive of the type of information that can be tested by the customary measurement techniques. The construction of devices for the measurement of information is discussed and illustrated in Chapter XI.

Quality. The quality of a project depends on how well the various tool operations have been executed, assuming, of course, a constant quality of material. It is the composite result of the type of material used and the skill with which it was worked. Quality involves such factors as squareness, roundness, finish, fasteners, exactness of dimensions, accurate placement of parts, etc. To secure a reliable measure of quality, it is usually necessary to employ a performance test. The pupil or pupils being tested make a standard project which gives samples of their work with different tools, instruments, or machines. The results obtained are then rated on suitable scales of quality, and thus the teacher is able to obtain a reasonably objective estimate of quality of work. A further discussion of the measurement of quality is given in Chapters XI and XIII.

Quality may be scored by physical measurement, by the use of scales, and by general observations of the student's procedure and product. The following are examples of measurable qualities from several industrial education subjects. The examples reported in Table 17 are taken from courses of study, committee reports, and the results of job analysis. They are not given as complete but should prove

suggestive for test workers and teachers who may desire to construct performance tests.

Technique. In general, the individual who manipulates the tools, machines, or instruments in the shop or at his desk in the most direct and efficient manner has the best technique. A pupil with good technique can improve his skill to the maximum of his ability by practice. Before technique can be measured it is necessary to know just what constitutes the best technique for using different tools and machines

TABLE 17

EXAMPLES OF OPERATIONS THAT DETERMINE QUALITY IN INDUSTRIAL EDUCATION SUBJECTS

Subject	Operations
1. Woodwork	Planing, joining, nailing, screwing, sawing, measuring, shaving, scraping, boring, champfering, chiseling, gouging, filing, gluing.
2. Auto mechanics	Tighten bolts, install cotter pins, solder, measure in thousandths of an inch, fit bushing, fit bearings, fit rings, fit pistons, time motor, clean, grease.
3. Drawing	Closing corners, lettering, numbering, placement of drawing on page, measuring, neatness, dimensioning, circles, ellipse, irregular figures.
4. Electricity	Skinning wire, soldering, splicing, insulating, cutting wire, bending and straightening wire, installing conduit, installing binding posts, switch installation, drilling metal, concrete and brick, boring wood.
5. Home mechanics	Soldering, cutting wire, splicing, insulating, skinning wire, attaching wire to binding posts; drilling metal, brick, and concrete; planing, boring wood; chiseling, nailing, screwing, sawing, cutting pipe, tightening pipe joints, tightening belts, splicing belts, filing; applying varnish, enamel, and paint.

for different purposes. To measure technique, therefore, it is necessary to emphasize the best shop procedures for the tools used in the course and then check the technique of the pupil being tested through observation. Tests of this type are especially helpful in diagnosing pupil difficulties in the manipulative phases of the course of instruction. Tests of technique should prove of special value in certain types of trade- and continuation-school classes.

Speed. Speed is determined by the time it takes to do a job with the quality held at a standard. Unless the quality is held constant,

the time required to do a piece of work is not a suitable basis for determining speed. Speed has considerable importance in trade courses, but in cultural courses of the junior-high-school type it is of much less significance. If a student has good technique he can develop his maximum speed through practice. At the present time the authors do not know of any suitable tests of speed which are available for general use in industrial education.

Reading Technical Symbols. It is necessary to read drawings and interpret various types of symbols in industrial education. In life and in school it is often of more practical use to be able to read symbols than to make them. In a trade course where the objective is strictly vocational, both the making and reading of symbols may be of importance. In order to get an adequate measure of a pupil's ability to read drawings or symbols, it is necessary to construct objective tests which give the pupil an opportunity to read enough drawings to determine his ability in this respect. No adequate tests of this type have appeared thus far for general use, but sufficient work has been done to demonstrate their feasibility.

Reading. Teachers of industrial education should quickly discover the reading abilities and limitations of their students. Those who are defective in reading should be given remedial help designed to overcome the difficulty, rather than allow them to be penalized by a poor mark because they are unable to understand the printed instructions. Since written instruction sheets are coming into common use in industrial education subjects, reading is especially important. A number of excellent tests of reading ability are available.

Spelling. More written work is found in certain phases of shop work than formerly, and the shop teacher should assume his share of the responsibility for the elimination of spelling difficulties. Certain technical terms should be mastered by the pupil so that he has no difficulty in pronouncing or spelling them correctly. Such training should be included as part of the regular course. Considerable research has been done in spelling, and many good spelling tests are available for determining spelling difficulties. Suitable spelling tests can easily be constructed for the purpose of measuring the pupils' mastery of the technical words used in the respective courses. Industrial education teachers will do well to make the pupil's spelling ability entirely distinct from achievement in the course.

Mathematics. Many shop courses involve related mathematics. The student usually has some mathematical background when he comes into the shop, but often many details have been forgotten. Accordingly, there are ordinarily wide differences in the abilities of

the pupils in the class. Some may be able to do the related mathematics whereas others may need remedial instruction. A few tests in shop mathematics have appeared, but they have been more general than specific. There is still a need for tests which treat the mathematics needed for special industrial education subjects. Some industrial subjects in which related diagnostic tests in mathematics may be used effectively are printing, electricity, woodwork, auto mechanics, sheet metal, and machine shop.

Appreciation of Industrial Products. One of the major objectives of junior-high-school industrial arts is the development of the consumer's appreciation of industrial products. The majority of people are more frequently buyers of industrial commodities than they are the actual producers. This means that an appreciation of the products of industry and the trades is important and should be measured. Little usable material is now available for the measurement of this factor. It, therefore, presents an unusual challenge to test workers in industrial education. Psychologically, the selection of an article involves the making of a judgment based on the composite of several variable factors. For example, if it is desired to select a kitchen chair, the following factors might come into consideration: material, cost, utility, designs, weight, strength, and finish. It is obvious that a consumer must have a knowledge of the qualities of the commodity, and some experience in evaluating them, before an adequate selection can be made. At the present time a rating scale seems the most satisfactory means of developing and measuring consumers' appreciation. Specific suggestions on the construction of such rating scales are given in Chapter XII.

Planning. Planning involves the ability to map out a direct and effective method for doing a job. It is generally conceded that, before a workman can plan the best method for doing a job, he must have an understanding of the factors involved. Ability to plan is now generally given as one of the desirable objectives to be achieved in industrial education courses. Individual differences are as great in ability to plan a project as in ability in other directions. Some of this difference in ability to plan is due to lack of training and some to native capacity. The habit of attacking problems in an orderly manner is a valuable one in any type of occupation, and industrial education work offers many opportunities for its exercise. Here again no suitable tests of planning have appeared, although they would be very useful in measuring success in shop courses. Some evidence of a student's ability to plan, as it relates to a single subject, may be secured by giving a number of situations which require the formula-

tion of a plan. The plans proposed by the students may then be compared with an ideal solution and with proposals of other pupils having similar backgrounds.

Language. The industrial education teacher needs to give attention to the language difficulties of his students in order to be of greater service to them in developing desirable oral and written language habits. The language of industrial education courses involves essentially the same principles as those governing correct usage in any subject. This enables the industrial education teacher to make use of available diagnostic tests for determining language difficulties. Here, as in spelling, teachers of industrial education should distinguish between achievement in industrial education courses and development in the use of language. A pupil's achievement in an industrial education subject should not be penalized because of language errors. Achievement in a course in woodworking is one thing, and the mistakes a student makes in language are quite another. The wise and sympathetic teacher will give the pupil special help designed to overcome his language difficulties, rather than penalize him by lowering his mark in industrial education achievement.

Inventiveness. Psychologically, inventiveness is similar to planning, in that it involves the formulation of a plan or series of plans. It obviously requires a higher type of mental ability, even to the extent of demanding abstract thinking with a dash of constructive imagination thrown in. Every shop teacher has met the boy who believes he has the solution of perpetual motion, or who knows he can improve available shop equipment by making certain changes in the machines. The authors have had several experiences in which boys sanding wood by hand near the wood lathes have conceived the idea of making a cylindrical sander by using the lathe. Probably not one of these boys had ever seen a power sander, and would have been greatly surprised as well as disappointed to learn that their invention had been used for several generations. So far, no adequate test of inventiveness has been developed.

Personality Traits. In recent years much consideration has been given to the significance of personality or character traits. The American Vocational Association committee on standards of accomplishment in industrial arts [2] refers to this phase of the work as "what you should be," and suggests the following traits as being worthy of development because of their recognition as essential to success in life: industry, cooperation, consideration of others, self-reliance, and readi-

[2] Standards of Attainment in Industrial Arts Teaching, *Bulletin of the American Vocational Association,* New York, December 12, 1931, pp. 21.

ness to assume responsibility. Character development is certainly an important phase of education for a democracy. Industrial education teachers have many opportunities to develop these traits in their students. Ordinarily, personality or character traits are measured by means of a rating scale. Several such scales have been developed for general use, but no published tests have so far appeared for rating pupils in the shop.

Mechanical Aptitude. Mechanical aptitude may be thought of as the capacity of a pupil to deal successfully with mechanical devices. Mechanical aptitude is now generally recognized as a measurable quality. It varies considerably among individuals and, in general, has a low correlation with intelligence scores. A knowledge of a pupil's mechanical ability is of value in assigning projects and in giving guidance suggestions for industrial vocations. Considerable research has been done on mechanical aptitude, and several good tests are available.

Intelligence. General intelligence is commonly considered the ability of an individual to learn. A knowledge of a pupil's general intelligence is of considerable importance in teaching and in educational guidance. To date, more than two hundred tests of intelligence have been published. The majority of these have received little consideration because of the lack of adequate validation, or the unreliable measures resulting from their use.

SUMMARY

Industrial education teachers need precise knowledge about their pupils as an aid in teaching, rating, and guidance. Fifteen important and measurable factors in industrial education are: information, quality, technique, speed, reading technical symbols, reading, spelling, mathematics, appreciation of industrial products, planning, language, inventiveness, personality traits, mechanical aptitude, and intelligence. Many of these measurable factors are a distinct challenge to test workers and teachers in industrial education. In general, the measurable factors of industrial education can be tested most effectively when they are measured individually or in a separate division of a test.

SUMMARY EXERCISES FOR DISCUSSION

1. Define and illustrate each of the measurable factors listed in Table 16.
2. Outline a plan for measuring objectively a pupil's ability to plan an attack on an industrial education problem.
3. State three or more objective questions or exercises designed to measure inventiveness.

4. Show how reading ability is an important factor in the measurement of achievement in industrial education.
5. Make a list of the technical words that pupils in industrial education course should be able to spell.

SELECTED REFERENCES

LEAVITT, F. M., "Standardized Measurements in the Field of Industrial Arts," *Industrial Arts Magazine,* Vol. 8: 132, April, 1919.

NEWKIRK, L. V., and STODDARD, GEO. D., *The General Shop.* Peoria, Illinois: The Manual Arts Press, 1929.

Report of Committee on Standards of Attainment in Industrial Arts, *Bulletin of the American Vocational Association, Industrial Arts Section,* December, 1931. pp. 20-31.

SWOPE, AMMON, "How to Construct Objective Tests in Industrial Subjects," *Industrial Education Magazine,* Vol. 30: 7-9, No. 1, July, 1928.

CHAPTER VI

ADMINISTERING INDUSTRIAL EDUCATION TESTS

29. Responsibility for Giving and Scoring the Tests.

The matter of determining the responsibility for giving and scoring educational tests rests chiefly upon the function the tests are expected to perform. If the tests are of the narrow-function type, closely paralleling the course of study taught by the teacher, they should undoubtedly be given by the teacher himself. If they are designed for survey purposes, or if the results are to be used for experimental, supervisory, or research purposes, they should probably be given by some one representing the administrative office of the school. Since these latter uses of the modern educational test are by far in the minority in most school systems, it is obvious that most of the classroom testing will be done by the classroom teacher.

An excellent generalization for determining the responsibility for the administration of educational tests may be stated as follows: Whenever the test results are of a type to provide the teacher with a reliable and valid basis for the discovery of individual pupil difficulties in learning or achievement, they should so far as possible be administered and interpreted by the teacher himself; otherwise, they should be administered by some other school official or a disinterested party. The single important exception to this general policy is the individual mental test, which, of course, should be given by a trained and experienced examiner other than the classroom teacher.

Properly selected tests for classroom use will contain so much valuable information that the teacher, in most cases, will be robbed of a rich opportunity to learn about his pupils and his own instructional efficiency if he does not insist upon his right to score the papers himself. This is particularly true of industrial arts subjects, in which the instructor will be mainly interested in the pupil's mastery of content. Teachers should regard the scoring of educational tests, whether they be those selected and used by themselves with their own classes or superimposed tests, as a personal responsibility and as an opportunity for securing significant information which should distinctly improve their teaching practice.

30. When to Give Tests.

The matter of when to use an educational test in the classroom, like the location of the responsibility for giving it, is determined almost entirely by the function the test is to perform. In the period of test development when the tests were not so numerous and lacked sufficient reliability for individual pupil analysis, the common practice was to administer a test at the end of the school term. This was adequate to give a general picture of the end-product of instruction, but it failed to accomplish two very important things from an educational point of view. In the first place, assuming that one of the very important functions of school training is to bring about changes in the quality of pupil response or the level of mastery, such a procedure gives no basis for such an evaluation. Progress cannot be determined by end-of-the-year measurement alone. In the second place, any weaknesses revealed by this end-of-the-year measurement are brought out too late to permit anything to be done about them. Remedial and corrective instruction under these conditions of measurement is impossible. Accordingly, many teachers are now making use of similar forms of tests at the beginning of the year as a check on initial status, and again at the end of the year as a measure of the year's accomplishment. This type of measurement permits an evaluation of initial status and the relative efficiency of instruction during the year, as well as presenting a fairly accurate picture of pupil growth in mastery during the period.

The development of extensive and detailed tests based upon a much more critical analysis of the different fields of instruction gave rise to a more refined idea of the use of tests. Clearly, the use of the test at the beginning of the term was justified mainly by the fact that it made it possible for educational progress or improvement to be evaluated. The end-of-the-year test was justified largely on the same basis, for it certainly did not provide any adequate basis for corrective instruction since the pupils involved were likely to be out of the hands of the teachers by the time the tests themselves were corrected and interpreted. The next logical step in the use of tests, therefore, was to construct many tests measuring a rather limited unit of instructional material. These tests made it possible for the teacher to make an immediate check-up on the efficiency of instruction as soon as the teaching of a particular unit of subject-matter was completed. The fact that these tests were each standardized as of the end of the particular instructional period involved made them especially useful to the teacher as the basis for organizing corrective and preventive work.

The narrow-function unit-type tests have been slow to develop in

the industrial education field although several useful contributions have been made. For the most part these tests are not standardized, nor has their reliability of measurement been critically checked. However, they present useful material and should point the way to additional contributions in this field. The validity of the tests has been determined with much more care than has the reliability. It is probably safe to assume, however, that the reliability, low as it may be in certain cases, is much better than the teacher's subjective judgment.

The *Nash-Van Duzee Instructional Review Test in Mechanical Drawing*[1] is a good example of a series of short tests based on a carefully validated group of instructional divisions.

These tests include the following subject-matter units which were selected after an analysis of textbooks, courses of study, and drawing teachers' judgments: drawing instruments and their use, terms and definitions, lettering, orthographic drawings, working drawing, sections, graphs, inking technique, construction problems, developments, materials, screw threads, conventions, fastenings, pictorial drawing, architectural drawing, gears, cams, detailed drawing, and assembly drawing.

The *Hunter Shop Tests*[2] also have value in measuring a number of instructional units in different phases of industrial education. However, the validity has not been critically determined by a check of representative courses of study and cannot be used for the detailed analysis that is possible with the *Nash-Van Duzee Instructional Review Tests in Mechanical Drawing.*

It is thus clear that the program of test administration in this, as in other instructional fields, is largely conditioned by the use to be made of the results. In general, if the function is mainly that of a survey for the purpose of comparing school with school, or class with class, the use of the test at the end of the school year may be adequate. If evaluation of pupil development in mastery or achievement is the object, then cross-sections at the beginning and at the end of the instructional period should be taken. If an objective basis for immediate pupil adjustment through remedial and corrective instruction is the major purpose, then the narrow-function unit-type tests must be used systematically throughout the year. Naturally, this type of testing program is fairly expensive in terms of time, pupil-teacher effort, and financial outlay. However, there are serious doubts whether in the last analysis any other type of testing program can be justified.

[1] Nash, Harry B., and Van Duzee, R. R., *Instructional Review Tests in Mechanical Drawing*, Bruce Publishing Company, Milwaukee, Wisconsin, 1930.

[2] Hunter, Wm. H., *Hunter Shop Tests*, Manual Arts Press, Peoria, Illinois.

The teacher has a right to expect a tangible return in the form of supervisory suggestions and remedial helps for his class for any time spent in taking, giving, scoring, or interpreting educational tests.

31. Controlling the Variables in Testing.

Most modern tests on the informational aspects of industrial education are constructed in such ways that almost any shop teacher who is reasonably skillful in maintaining the discipline of his class and who will follow the directions accompanying the tests can administer them without difficulty. It is always desirable, of course, that the teacher should become very familiar with the examiner's manual before attempting to give any kind of test. If the examiner is inexperienced in the giving of tests he should try the test out on someone before giving it to his class. If this is impossible, the test itself and the directions for administering it should be read through several times before attempting to give it to a class.

The following general suggestions may be useful to the individual not widely experienced in the administration of tests:

1. Before beginning the tests have the desks cleared and see that each pupil is provided with one or more pencils. Have a number of extra pencils available for emergencies.
2. The room should be quiet throughout the tests. Require strict attention to the directions, and see that the pupils follow your commands at once. If the group tested is large, additional proctors may be necessary. They should move quietly about the room and see that all pupils get started correctly and together.
3. The examiner (and proctors) should pass down the aisles and place a test booklet on the desk of each pupil with the cover page (page 1) facing the pupil. If the tests are in mimeographed form, place the folders face down and instruct pupils to leave them in that position until they are told to fill in the blanks at the top of the first page.
4. All directions to the pupils should be given carefully in a tone which carries proper emphasis and suggests authority. The voice should be just loud enough to be heard in all parts of the room used for testing.
5. Follow the directions of each test strictly, and adhere rigidly to the time limits. A stopwatch is highly desirable for timing the tests.

6. See that all pupils start and stop instantly upon the signal. Students should be instructed that, should they finish a test before time is called, they may go over their work and look for mistakes.

32. Administering Manipulative Tests.

Manipulative or performance tests present problems of administration which differ in certain respects from the objective pencil-and-paper tests of information. In a performance test the pupil modifies materials with tools, makes a drawing with instruments, or applies a coating to a surface. The performance test is a recognition test in which the pupil selects the tools or instruments which are already available and turns out a product which can be rated objectively and compared with other, similar products.

Like tests of the paper-and-pencil variety, manipulative tests usually have prepared sets of directions which must be studied by the teacher and carefully followed. It is also advisable for the teacher to practice giving a performance test to individual pupils or to a small group before attempting to give it to an entire class. Manipulative tests require very careful supervision in order that the resulting product may be uniform and thus lend itself to objective rating.

The authors have found the following points helpful in administering manipulative tests:

1. Read aloud and distinctly the directions to pupils while the class follows silently.
2. Answer all questions about the directions before the test is begun.
3. Show the pupils a completed test-product, and, if they care to, let them examine it.
4. If there are no further questions, say, "Get ready. First tool or instrument up. Go." (Record time.)
5. During the examination answer any questions about the steps in the procedure by reading the steps in question with the pupil.
6. Observe the pupils as they work to make certain they are all doing the steps in the correct order.
7. Make certain that the proper tool or instrument is used where indicated, but do not tell the pupils how to use the tool.
8. Help any pupil having trouble in reading working drawings, but do not make measurements on his problem for him.

Other factors which the authors have found important to consider in administering a manipulative test are: (1) the condition and placement of tools or instruments, (2) quality of materials, (3) lighting, and (4) convenience and comfort of place to work. The tools used must be of suitable size, properly adjusted, and uniformly sharp for each student taking the test. The materials to be used should be of uniform quality and free from defects. If one pupil has a piece of knotty oak and another one a piece of clear gum-wood, it is obvious that the two pupils are not working under comparable conditions and have not the same chance to obtain good results. It is conceivable that the pupil with the knotty oak might do better work and yet get a lower rating. Proper lighting is also very important in a performance test, because it is necessary for the pupils to get clear images of their work. A suitable light for a manipulative test is 12 to 15 foot-candles of illumination on the bench top or drafting-board surface. It is also well to be certain that all pupils taking a manipulative test have normal eyesight. The benches or drafting boards should be of proper height for the individuals taking the test. It too frequently happens that extremely tall or short pupils are handicapped by too low or too high a working surface.

33. Scoring Manipulative Tests.

Manipulative tests are scored by physical measurements, rating scales, and observation of experts. It is obvious from this statement that the products of manipulative tests cannot be scored as objectively as the pencil-and-paper tests. However, the scoring is much more objective than the teacher's subjective judgment, and usually reasonable objectivity can be secured. The most objective scores are obtained where physical measurements can be used. If a pupil cuts a board 13½ inches long when it should have been 14 inches, the board can be measured and the amount of the error determined in fractions of an inch. This is as objective as the response to a true-false question. The rating of a soldered joint, a rope splice, a glued joint, the setting of a flat-head screw, or the tying of an underwriter's knot are modifications of materials, but they are the result of an almost limitless number of small variables which produce a quality of workmanship that is not easily measured by instruments or readily judged by experts without means of comparsion. Since this is true, joints, splices, lettering, etc., can be rated most objectively by comparing them with samples of known quality. This rating process is referred to as using a rating scale. The construction and use of rating scales are discussed in detail in Chapter XII.

There are other characteristics of the results of manipulative tests which do not lend themselves to physical measurement, but which can be rated quite objectively by experienced observers. Examples of these are letters transposed in printing, or loose wires in electrical circuits. The important thing in achieving objectivity in scoring manipulative tests is to use the techniques of physical measurement, quality scales, and experienced observers in rating those results to which they are adapted. If this is done the objectivity of the scoring of manipulative tests may be made quite satisfactory.

34. Responsibility for Training in the Use of Scales.

The use of scales for the more or less objective rating of qualities or products has been pointed out as an important phase of measurement in industrial education. In fact, many outcomes of shop work are measurable in no other way than by rating scales. For example, the quality of a soldered joint in metal work or of a mitered joint in woodwork is not readily evaluated objectively except through the use of a scale. Experience with such rating scales makes it apparent, however, that reliable results cannot be obtained from their use by untrained and inexperienced judges. Brief courses of training in the use of the scales result in distinctly reducing the unreliability of measurement resulting from the subjective factors. Classroom teachers can be trained in the use of handwriting scales, freehand drawing scales, composition scales, and doubtless many other kinds of scales, to the point where the average error or deviation from a known quality rating will not exceed five points on a hundred-unit scale. This is probably not a serious inaccuracy in such measurement. It may be inferred therefore that similar training periods must be provided for industrial education teachers who are desirous of using rating scales in this field. Increased reliability of measurement may be expected as a result of such training.

The head of the department of industrial education in a large high school may well assume the responsibility for giving his teachers a brief course of training in the use of rating scales. Typical samples taken from the chosen field may be used for this practice. Preferably, samples representing a wide range of quality should be chosen. If the samples are selected from the products of a class and the true quality scores of the samples are not known, the average ratings given by a group of six or seven teachers may be taken as the basis for adjustment. Judges whose ratings deviate most widely from the composite or the true values should be asked to rerate their samples, making certain conscious adjustments in their mental standards of

quality until they conform quite closely to the standards of the group. Considerable experience in working with training-groups in the use of such scales indicates that certain individuals readily adjust their ideas or standards of quality to those of the scale. There is some slight evidence that such individuals, being gifted with greater discriminative power, are usually found among the more able groups of teachers. For such individuals a brief period of training is adequate. For the average teacher, inexperienced in the use of such scales, as many as two or three hourly periods of drill in the rating of the selected specimens may be necessary before a satisfactory level of accuracy of measurement is reached.

35. Scoring Pencil-and-Paper Tests.

One of the important distinguishing features which characterizes objective paper-and-pencil tests is their very objectivity. Objectivity in a test implies little or no variability in the acceptable answers. Objective tests should be scored in exact accordance with the scoring key. The directions should be followed rigorously and the tests scored exactly according to instructions, even though they may run counter to the user's best judgment. Unless this care in scoring the tests is taken, it is impossible and improper to make comparisons of the test results with the norms or standards which have been derived under controlled conditions. Errors in scoring and transcribing test scores are best eliminated by rechecking all such work and by performing all related calculations at least twice. Special care should be taken where the results are to be used for experimental purposes or for individual pupil analysis.

The remaining phases of the administration of tests in the classroom are essentially statistical and interpretational in character and as such are reserved for discussion in Chapters XIV and XV.

SUMMARY

The matter of determining the responsibility for the giving and scoring of educational tests rests to a large degree upon the use to be made of the test results. The questions of when to use an educational test and what kind of a test to use are answered almost entirely by the function the test is to perform.

Tests in the industrial education field are broadly divided into (1) tests of information and (2) tests of performance. Information tests are commonly of the paper-and-pencil variety, calling for evidence of a mental reaction. Performance tests are manipulative and constructive in character, calling upon the student to apply tools and skills

to materials, and produce tangible objects of varying quality in accordance with certain definite specifications. Conditions under which tests of both types are administered must be carefully controlled if the results are to be meaningful.

SUMMARY EXERCISES FOR DISCUSSION

1. What should be the classroom teacher's responsibility for the administration and interpretation of industrial education tests?
2. What factors determine primarily what tests to give and when to give them?
3. In what specific points does the administration of the objective test of the paper-and-pencil type differ from that of the manipulative type of industrial arts test?
4. How may the scoring of manipulative tests be objectified?
5. What does the evidence show regarding the influence of rating scales for shop products on the reliability of marks assigned? Does training in the use of such scales appear to pay?

SELECTED REFERENCES

Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests.* New York: Longmans, Green and Company, 1935.

Gregory, C. A., *Fundamentals of Educational Measurements,* Boston: Houghton Mifflin Company, 1923.

McCall, W. A., *How to Measure in Education.* New York: The Macmillan Company, 1922.

Odell, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

Ruch, G. M., and Stoddard, Geo. D., *Tests and Measurements in High School Instruction.* Yonkers, New York: World Book Company, 1927.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

Trabue, M. R., *Measuring Results in Education.* New York: American Book Company, 1924.

Wilson, G. M., and Hoke, K. J., *How to Measure* (Revised). New York: The Macmillan Company, 1928.

Woody, Clifford, and Sangren, Paul V., *Administration of the Testing Program.* Yonkers, New York: World Book Company, 1933.

CHAPTER VII

INDUSTRIAL EDUCATION ACHIEVEMENT TESTS

Selected tests for certain fields of industrial education are described and evaluated in this chapter. In general, the tests named have been published and widely distributed during the past few years. Many are quite satisfactory and are in most respects the equal of good tests in other fields, but quite a number are pioneer efforts and are included more because they are suggestive for future development along more scientific lines than for their present merit. Many more carefully prepared standardized tests in industrial education are needed before the field will be as well covered as other fields of instruction. No attempt has been made to include all the available tests, but tests which seem to have special values for industrial education teachers have been selected from the different instructional fields.

36. Achievement Tests in Industrial Education.

In order to measure achievement in industrial education, it is necessary to measure information and ability to perform tasks involving the use of tools, machines, and materials. Ability to perform a task does correlate with knowledge, but the relationship does not seem to be sufficiently close to warrant the use of the pencil-and-paper-type test to measure all types of achievement in industrial education. For example, a pupil may know how to do a job and be able to do it if given the opportunity, and yet make a poor score on a pencil-and-paper test because he does not know the technical vocabulary. Another pupil may know the procedure and the vocabulary but lack the tool skill necessary for the execution of the project.

The fact that the work in industrial education is not well standardized from school to school has been pointed out by many test workers as an insurmountable obstacle in the way of the construction of standardized industrial education tests, and it has been contended that further test construction should wait until the work is more definitely standardized. On the surface this seems logical, but in fact it has little foundation because, in other fields of instruction, the research work needed to validate a test has been one of the chief influences tending to standardize curricular content and establish levels of ac-

complishment. In industrial education, as in other subjects, it is necessary to make a careful study of teaching practice, textbooks, courses of study, committee reports, and, in many cases, to make extensive analysis of the subject to be tested. All these studies lead toward a better understanding of the content in any of the subjects selected for test construction. These validation studies have a marked influence on teaching practice because they are put in the form of a test and the teacher can determine in part whether or not his course is valid by comparing it with the items in the test and the median results obtained with those of other schools. It is obvious, therefore, that standardized tests of achievement with their attending validation studies are one of the strongest influences tending to define and set up standards of accomplishment in industrial education courses. It is of course difficult to establish norms that are of great value, but norms can be revised as the work becomes more uniform through the use of validation studies.

37. Standardized Industrial Arts Tests.

In this section four widely used standardized industrial arts tests are briefly described.

1. Nash-Van Duzee Woodwork Test 1, Scale A[1]

This is a test designed to measure the junior- and senior-high-school pupil's knowledge of processes, tools, materials, and information used in woodworking. Five different types of questions are used in the test with appropriate directions for each type. The test is printed in a neat, eleven-page booklet and is accompanied by a manual of directions, objective scoring key, and class record card. In general the test has been carefully constructed.

Validity. The validity of the test was based on an analysis of teaching content as obtained from courses of study, textbooks, surveys, reference books, and trade analyses. Apparently the validity of the test is satisfactory in the light of common practice.

Reliability. The coefficient of reliability based on 200 cases was found to be .86, using the chance-half method and employing the Spearman-Brown formula for estimation of the reliability of the whole test. The reliability of the separate divisions of the test on the same number of cases is given in Table 18. Although the reliability of this test is a little below the best academic tests of the same type, it is very satisfactory for measuring achievement for comparative purposes.

[1] Nash, Harry B., and Van Duzee, Roy R., *Woodwork Test 1, Scale A,* Bruce Publishing Company, Milwaukee, 1927.

Norms. Norms are reported on 3000 cases. They are given on the basis of semesters and number of minutes of instruction from the first semester of the junior high school through the first two semesters of

TABLE 18

RELIABILITY COEFFICIENTS FOR NASH-VAN DUZEE WOODWORK TEST

Part		Reliability Coefficient
I	A	.61
I	B	.80
I	Total	.85
II		.80
III		.94
IV		.88

high school. A few cases are also reported on training-school students. The norms are likewise given in the form of percentiles with a corresponding marking scale. The methods employed in securing and reporting these norms should be very suggestive to other test workers in industrial education.

2. NASH-VAN DUZEE WOODWORK TEST 1, SCALE B [2]

This is a test for the purpose of measuring the pupil's skill in manipulating hand woodworking tools. It is a companion test of *Test 1, Scale A,* which is a test of information rather than performance. The test is suitable for measuring manipulative achievement in junior- and senior-high-school hand woodwork. Nash and Van Duzee [3] state in the manual that "the test aims to measure the pupil's understanding of directions involving frequently used woodworking processes and procedures, the reading of a working drawing, the selection of proper tools to carry out the specified work and the ability to use the tools selected to do the required work."

Validity. The skills for the test were selected after analyzing courses of study, textbooks, problem books, and blueprints used in junior and senior high schools. The items selected are representative of general practice, but there is a question as to whether enough of each type of item is given to really sample the pupil's ability.

[2] Nash, Harry B., and Van Duzee, Roy R., *Woodwork Test 1, Scale B,* Bruce Publishing Company, Milwaukee, Wisconsin, 1928.

[3] Nash, Harry B., and Van Duzee, Roy R., *Manual of Directions Industrial Arts Test 1, Scale B,* Bruce Publishing Company, Milwaukee, Wisconsin.

Reliability. The reliability of this test is reported as varying from .60 to .80 with an average of about .73. This is higher than a teacher's subjective judgment but is probably too low for a first-class standardized test.

Norms. Median and percentile norms based on a few hundred cases are available.

3. Newkirk-Stoddard Home Mechanics Test [4]

The chief purpose of the *Newkirk-Stoddard Home Mechanics Test* is to measure in an objective and analytical manner the essential knowledge that the pupils should acquire from a well-organized course in home mechanics. The test is divided into two closely equivalent forms, A and B. Each form contains 36 jobs, comprising a test of half the outstanding jobs in home mechanics. It is divided into Forms A and B so that it will be easier to administer and will more nearly fit the various needs of home mechanics teachers.

Validity. Four criteria were used to establish the validity of the test:

1. Surveys to check the jobs on the basis of social utility.
2. Surveys of the actual teaching content of 75 representative schools.
3. Analysis of course of study and widely used commercial job sheets which were based on surveys.
4. Selection of jobs with procedures representative of a class of jobs rather than just a single job.

Scoring. The test is objective in its scoring. Each form has a printed key in which the correct responses for Part I are placed around the margins and the correct diagrams for Part II are reproduced on the back. This scoring key contains full directions for its use. The key sheet requires no cutting and is easy to use. No corrections for chance are required.

In Tables 19 and 20 are given statistical measures which indicate the consistency of measurement by the test.

[4] Newkirk, Louis V., and Stoddard, George D., *Newkirk-Stoddard Home Mechanics Test,* Bureau of Educational Research and Service, University of Iowa, Iowa City, 1928.

TABLE 19

RELIABILITY OF A SINGLE FORM (A OR B) 40 MINUTES' TESTING

Grade	No. in Sample	r		Standard Deviation		P.E. Score	
		Job	Point	Job	Point	Job	Point
7	50	.49	.86	2.5	26	1.2	6.6
8	50	.64	.89	3.0	29	1.2	6.5
9	50	.59	.85	3.2	24	1.4	6.3
7-9	150	.54	.86	2.9	27	1.3	6.8

TABLE 20

RELIABILITY OF BOTH FORMS (A + B) 80 MINUTES' TESTING

Grade	No. in Sample	r		Standard Deviation		P.E. Score	
		Job	Point	Job	Point	Job	Point
7	50	.66	.92	4.5	50	1.6	9.5
8	50	.78	.94	5.5	56	1.5	9.3
9	50	.75	.92	5.7	45	1.9	8.6
7-9	150	.70	.93	5.4	53	2.0	9.4

Norms. Table 21 recapitulates preliminary norms obtained from grades 7, 8, and 9.

TABLE 21

NORMS, FORMS A AND B, MAY TESTING ($N = 396$)

	Jobs		Points	
	Form A	Form B	Form A	Form B
Mean	5.2	5.6	74.7	74.6
Median	4.8	5.2	77.9	73.0
Upper quartile	7.0	7.6	95.8	96.3
Lower quartile	3.2	3.1	60.5	54.5
Standard deviation	2.9	3.4	25.1	27.9

4. Wells-Laubach Industrial Arts Tests [5]

Tests in woodwork, printing, machine shop, and mechanical drawing are included in this series. All the tests are of the pencil-and-paper type and with the exception of one in mechanical drawing are made up of 100 true-false statements. Twenty-five minutes is the working time for the woodwork and printing tests, 20 minutes for the machine shop, and 30 minutes for mechanical drawing.

Validity. The content of the tests parallels teaching practice in a general way but no scientific means of determining validity is reported.

Reliability and Norms. No statistical data are given on the tests, but tentative norms based on the median accomplishment in about 1000 cases are reported. The tentative norms are given on the basis of four semesters for each of the four tests.

38. Non-Standardized Industrial Arts Tests.

1. Hunter Shop Tests, Series 1 and 2 [6]

Hunter has developed 32 short objective tests. Each test includes 25 objective questions on the particular subject measured. The following is a list of these objective tests according to the subjects or parts of subjects tested. They are of the pencil-and-paper type.

Woodwork

W–1	Tools test
W–2	Fastenings test
W–3	Comprehension test
W–4	Trade names test
W–5	Reading test for rule or scale
W–6	True-false test
W–7	Completion test
W–8	Building parts test for carpentry
W–9	Board measure test
W–10	Multiple-choice test for wood and lumber
W–11	Reading test for framing square
W–12	Objective test for wood finishing
W–13	Multiple-choice test for carpentry
W–14	Multiple-choice test for pattern makers

Mechanical Drawing

MD–1	Reading test
MD–2	Missing-line test
MD–3	Lettering test
MD–4	True-false test

[5] The Manual Arts Press, Peoria, Illinois, 1928.
[6] The Manual Arts Press, Peoria, Illinois, 1927.

MACHINE SHOP

MS-1	Tool test
MS-2	Comprehension test
MS-3	True-false test
MS-4	Micrometer reading test
MS-5	Multiple-choice test

ELECTRIC SHOP

E-1	Symbols test
E-2	Objective test

AUTOMOBILE MECHANICS

AM-1	Parts test
AM-2	Multiple-choice test

PRINTING

P-1	Completion test

RELATED SUBJECTS

G-1	Shop English
G-2	Shop mathematics
G-3	Shop arithmetic
G-4	Geometry

Validity. The validity has been determined in a general way from teaching experience, pooled judgment, and analysis of textbooks and courses of study.

Reliability and Norms. The reliability is not reported, and the tests are not standardized. They are too short to have very high reliability individually, but in using a battery of the woodworking tests the authors have found coefficients of reliability as high as .85.

39. Mechanical Drawing Tests.

Four tests in mechanical drawing are described in this section.

1. BADGER STANDARD TEST IN FUNDAMENTAL MECHANICAL DRAWING, TESTS 1, 2, 3 [7]

The author states that "these are tests of what the pupil knows about the phases of drawing covered rather than a test of his drawing ability measured in terms of neatness, accuracy, lettering and so forth." The test includes 145 exercises of the multiple-choice type. The form has been varied to fit the types of content tested. There are three tests. The first deals with knowledge relating to the use of instruments, line work, dimensioning, and lettering; the second tests knowl-

[7] Public School Publishing Company, Bloomington, Illinois, 1929.

edge of projection and includes sections and auxiliary views; and the third measures knowledge of pictorial drawing, isometric, cabinet, and oblique. The test does not have time limits, but the directions suggest that the tests be collected after all but two or three of the slowest pupils have finished. The validity and reliability of the test are not given in detail.

2. Castle Mechanical Drawing Test [8]

This test is divided into five subtests. Subtest 1 requires the pupil to identify similar parts of an object in top and side views by matching corresponding numbers and letters. Subtest 2 deals with dimensions; 3, with geometric terms; 4, with pencil technique; 5, with inking. The working time for the test is 41 minutes. The first three parts are objective in scoring, but the last two depend to some extent on the teacher's subjective judgment, although the scorer is provided with letter rating scales and six points are mentioned for rating the drawing.

Validity. No very definite statement is given as to validity, but it is based on analysis of instructional materials and a long teaching experience in mechanical drawing.

Reliability and Norms. The coefficient of reliability is not reported, and norms have not been established for the test. It is not standardized but should prove useful for measurement of drawing achievement in the same manner that a teacher-made objective test would be used.

3. Fischer Mechanical Drawing Tests, Parts I and II [9]

Part I of this test covers the technical information necessary in drawing. No instruments other than a pencil are needed. Part II is a performance test and requires the use of drawing instruments. Either test can be given in a 45-minute drawing period. Part I is composed of four subtests and Part II of three subtests. Both parts of the test should be given since it is desirable to test information and performance. Parts I and II are not equivalent forms but are divisions of the same test. The test has considerable diagnostic value as it enables the teacher to see where the pupils have succeeded and where they have failed. The problems in the test are rated according to difficulty. The test includes a manual, scoring key, and a class record sheet.

[8] The Manual Arts Press, Peoria, Illinois, 1928.

[9] The Bruce Publishing Company, Milwaukee, Wisconsin, 1929.

Validity. The claim for validity is based on analysis of textbooks, blueprints, courses of study, and in addition on a survey of 100 schools to find out what was being taught, time being devoted to drawing, etc. This material was tabulated under five major divisions as follows: use of instruments, lettering, projection drawing, geometric constructions, and pictorial representations. The content included in the test was carefully validated on the basis of teaching practice and represents good workmanship.

Reliability. The coefficient of reliability was determined by giving the same test twice to 150 sophomores in high school. This resulted in a correlation coefficient of .79. This is quite low for a standardized test.

Norms. Median-score graphs are given which indicate the median score for all schools as represented by scores from 2500 students. The norms or medians of accomplishment are classified on the basis of minutes of instruction. The author also suggests means of using bar diagrams and the use of test scores as a partial means of assigning marks.

4. Nash-Van Duzee Industrial Arts Test, Test II, Mechanical Drawing [10]

The test is designed to measure objectively performance in drawing as well as information about mechanical drawing. The test is suitable for use in both the junior and senior high school. The test is divided into Part I and Part II and is available in two closely equivalent forms. Forms I and II were equalized on the basis of the results obtained from 500 mechanical drawing pupils in the ninth and tenth grades. A manual of directions, objective scoring key, and a class record sheet are provided. The scoring key also includes a scale for the rating of ability to letter. Part I of either form can be written in the ordinary classroom with a pencil, but Part II requires the use of mechanical drawing instruments.

Validity and Construction. The claims for validity are based on the analysis of textbooks, courses of study and reference books, and a rating of the analysis by several hundred persons interested in teaching mechanical drawing.

Reliability. The reliability of the test was found to be .87 with 203 fifth- and sixth-semester pupils: Statistically, this is quite a satisfactory reliability since it was determined by correlating Form I with Form II.

[10] The Bruce Publishing Company, Milwaukee, Wisconsin.

Norms. Median and percentile norms indicating accomplishment by semesters and minutes based on 2500 cases are given for the junior high school and the first two years of high school. A suggestive scale for converting percentile norms into equivalent class marks is given in the table of percentile norms. Percentile curves for Forms I and II are used to show the approximate equivalence of the two forms of the test.

40. Trade Tests.

Trade tests are of value to industrial education teachers who teach vocational courses, and they are very suggestive to teachers of the general educational courses of the junior high school. Trade tests measure trade proficiency; they are valuable in selecting men who possess the information and skill necessary to succeed in a given trade and for measuring accomplishment in advanced vocational courses.

Chapman[11] has pointed out the significant distinctions between intelligence tests and trade tests. "While these two forms of the test, the mental test and the skill prediction test, both have a great sphere of usefulness in industry, it is very essential to precise thinking on the subject of industrial testing not to confuse these with the trade test proper. The trade test makes no pretense of measuring intelligence directly; it makes no attempt to measure the native endowment of the subject, with a view to predicting the degree of success to be expected as a result of training in a specific trade; the trade test furnishes a rating, in objective quantitative terms, of the degree of trade ability already possessed as a result of practice in the trade."

Trade tests present numerous difficulties in their construction and for that reason have not been entirely successful, although the better tests are decidedly superior to subjective judgments in selecting qualified tradesmen. One difficulty has been the lack of information of test workers about abilities, techniques, skills, and attitudes necessary for success in a given occupation to develop valid measures. Another difficulty is that trade tests are not always given under trade conditions, with the result that a man may succeed on the test but fail on the job. Trade tests are also expensive of time and money. Many of them are individual tests and require material and tools for the measurement of manipulative skills.

Trade tests of four general types—oral, picture, performance, and written group tests—were widely used in the army during the World War to select men who were proficient in the various trades. This

[11] Chapman, J. C., *Trade Tests*, Henry Holt and Company, New York, Chapter XI, p. 374, 1921.

procedure saved considerable time and money. Since the war many industries have employed trade tests in selecting applicants for positions. Tests of this type are used in vocational guidance. Trade tests have been greatly improved and modified during the past ten years and have been adapted to the needs of industry.

SUMMARY

Objective tests have appeared somewhat more slowly in industrial education than in certain other branches of instruction. Possibly this has been because of a lack of definiteness in the statements of the objectives of certain of the industrial courses.

Achievement in the industrial subjects is not entirely a matter of information. Ability to perform a task does correlate with knowledge about the task, but this relationship does not seem to be sufficiently high to warrant the exclusive use of pencil-and-paper tests for the measurement of achievement in the industrial subjects. Accordingly, performance as well as informational types of tests are needed in this field.

SUMMARY EXERCISES FOR DISCUSSION

1. Discuss the special limitations of paper-and-pencil tests in the industrial subjects.
2. Show how the fact that industrial education courses are not well standardized from school to school accounts for numerous difficulties in the construction of tests.
3. Select at least one test in the fields of woodworking, mechanical drawing, and shop work, and present the major values and limitations of each.

SELECTED REFERENCES

Badger, Alex J., *Standard Tests in the Fundamentals of Mechanical Drawing,* Tests 1, 2, and 3. Bloomington, Illinois: Public School Publishing Co., 1929.

Castle, Drew W., *Mechanical Drawing Test,* Peoria, Illinois: Manual Arts Press, 1927.

Castle, Drew W., "Mechanical Drawing Tests," *Vocational Education Magazine,* Vol. 2: 756-8, May, 1924.

Christy, Elmer W., *Mechanical Drawing Scale,* Peoria, Illinois: The Manual Arts Press, 1926.

Cleeton, Glen U., "Printing Tests for the Junior High School," *Industrial Arts and Vocational Education Magazine,* Vol. 19: 329, September, 1930.

Donson, George C., "A Machine Shop Test," *Industrial Arts and Vocational Education,* Vol. 20: 132-3, April, 1931.

Fischer, Ferdinand A. P., *Mechanical Drawing Tests,* Milwaukee, Wisconsin: Bruce Publishing Company, 1930.

Flaherty, Edward B., "Electrical Shop Tests for Grades 7, 8 and 9," *Industrial Arts and Vocational Education,* Vol. 19: 142, April, 1930.

FLAM, AUGUST, "First-Year Mechanical-Drawing Test," *Industrial Arts Magazine*, Vol. 17: 223, 336, June and September, 1928.

FLAM, AUGUST, "Second-Year Mechanical-Drawing Test," *Industrial Arts Magazine*, Vol. 17: 371, October, 1928.

FLAM, AUGUST, "First-Year Mechanical-Drawing Multiple-Choice Test," *Industrial Arts Magazine*, Vol. 17: 70, February and March, 1928.

FLAM, AUGUST, "Some Mechanical Drawing Tests," *Industrial Arts and Vocational Education*, Vol. 19: 150, April, 1930.

HJERTSTEDT, W. G., "Architectural Drawing Test," *Industrial Arts and Vocational Education*, Vol. 20: 141, April, 1931.

HJERTSTEDT, W. G., "Auto Mechanics Test," *Industrial Arts Magazine*, Vol. 18: 239-40, June, 1929.

HORNING, S. D., "Tests of Prognostic Value in Mechanical Drawing," *Industrial Education Magazine*, Vol. 29: 441-4, June, 1928.

HUNTER, WM. L., Shop Tests. Peoria, Illinois: Manual Arts Press, 1927-1930. Woodworking Tests, Mechanical Drawing Tests, Machine Shop Tests, Electrical Shop Tests, Auto Mechanics Tests, Printing Test, Related Subjects Tests.

JOHNSON, H. J., "Electrical Tests," *Industrial Arts and Vocational Education*, Vol. 20: 139-40, April, 1931.

KROLL, HARRY W., "A Mechanical Drawing Test," *Industrial Arts and Vocational Education*, Vol. 20: 127-9, April, 1931.

MURBACH, NELSON J., "Sample Woodworking Tests," *Industrial Arts and Vocational Education*, Vol. 20: 129-30, April, 1931.

NASH, HARRY B., and VAN DUZEE, ROY R., *Industrial Arts Tests*. Milwaukee, Wisconsin: Bruce Publishing Company. (Test I Scale A—*Woodwork Information*, 1928, Test I Scale B—*Woodwork Performance*, 1929, Test II *Mechanical Drawing*, Forms I and II, 1930. *Instructional Review Tests in Mechanical Drawing*, 1930.)

NEWKIRK, L. V., and STODDARD, G. D., *Home Mechanics Test*. Iowa City, Iowa: Bureau of Educational Research, 1928.

WEAVER, C. G., "Trade Tests: Their Construction, Use and Possibilities in Industry," *Industrial Arts Magazine*, Vol. 10: 163-6, May, 1921.

WELLS, G. K., *Shop Tests*. Peoria, Illinois: Manual Arts Press, 1929.

CHAPTER VIII

INTELLIGENCE AND APTITUDE TESTS IN INDUSTRIAL EDUCATION

I. MEASUREMENT OF INTELLIGENCE

41. Meaning of Intelligence.

The exact nature of intelligence is not well understood, but it is definitely known that individuals vary quite widely in mental ability and that within limits it can be measured. Authorities in the field of mental measurement are far from agreement as to what the term intelligence implies. Some consider that intelligence is best indicated by the ability of the individual to solve problems, to adapt himself to new situations. Others hold that the abilities to perceive with speed and accuracy, to associate symbols, to manipulate abstract concepts, and to reason, are the best evidences of intelligence. Facility in the use of language itself is considered by some to be one of the very significant evidences of intelligence. For the purposes of this discussion *intelligence will be considered as the capacity for learning, plus the informations, skills, and attitudes which the individual has gained from reacting to his environment.* This rather liberal conception of intelligence permits it to fit readily into its place in the educational program and also places in an acceptable light the majority of devices for the measurement of general mental ability.

42. Measurement of General Mental Ability.

Teachers and educators in general are aware that at least two related but different phases of intelligence must be taken into account in adequate mental measurement. Certain individuals react readily to abstract stimuli and thus are frequently rated as normal or even superior on the basis of mental tests in which abstract material predominates. Other types of individuals do not respond to abstractions but reveal unusual aptness in reacting to concrete and tangible material. Stenquist [1] has stated the case for this type of pupil most convincingly, and has presented a very useful supplement to the

[1] Stenquist, John L. "A Case for the Low I.Q.," *Journal of Educational Research,* Vol. 4: 241-54, November, 1921.

ordinary abstract type of mental measurement in his *Mechanical Aptitude Tests.*

It is true that much remains to be learned about intelligence and its measurement. There are those who argue that measures of intelligence should not be used because its exact nature is not known. This argument is no more valid than the statement that electricity should not be measured or used because its exact nature is not known. Intelligence, like electricity, can be measured to the distinct advantage of society if the results are properly used. Unquestionably the scores from mental tests do not reveal intelligence as exactly as the dials of an electric meter indicate the number of watts of electricity consumed. Nevertheless, a few reasonably reliable and valid measures of intelligence are available for general use. Odell [2] states that at least two hundred tests of mental ability have been constructed since the early work of Binet, and that approximately one hundred are still available for use.

43. Methods of Measuring Intelligence.

Intelligence tests are of two general types, individual mental examinations and group tests of mental ability. Individual mental examinations are thought to be considerably more valid, and because of the method of administering them they are probably more reliable than group tests. Individual examinations are expensive since they are given to only one subject at a time and since they should preferably be given only by a trained examiner well grounded in psychology. Much of the significance of the individual examination lies in the interpretations of the subject's reactions by the examiner as the stimuli are presented. Group tests are easy to administer, some being almost self-administering.

The problems of measuring intelligence commonly met by the teacher of industrial education can usually be handled satisfactorily by the use of carefully selected group tests. However, the group test results should almost certainly be supplemented by the individual mental examination for those having very high or very low scores and for problem cases. Where this is not possible, or where the problem is not extremely serious, the use of two or even three group mental tests is to be recommended. The average of the mental-age scores obtained from two or three group tests is a much more accurate measure than that ordinarily obtained from a single testing.

Since the problems of the industrial education teacher will ordi-

[2] Odell, C. W., *Educational Measurements in High School.* Chapter XV, pp. 391. The Century Company, New York, 1930.

narily be solved by the use of a good group test, and since most classroom teachers are not adequately trained or experienced in the use of the individual examination, a brief list of excellent group tests is presented for detailed consideration. A short description and evaluation of the most widely used individual mental examinations is given in a later section in this chapter.

44. Group Tests of Mental Ability.

The four group tests of mental ability selected for description and evaluation here are chosen from an extensive list of such tests. These tests are suitable for use in grades VII to XII, inclusive. Each test has been carefully validated and ranks high among such tests for the reliability of the test forms themselves as well as for the reliability of the age norms used as the basis of interpretation. These tests can be readily administered to a group of any number of pupils. The results are comparable within reasonable limits to those obtained on the individual examinations.

1. Kuhlmann-Anderson Intelligence Tests [3]

The *Kuhlmann-Anderson Intelligence Tests* are the result of more than ten years of careful research by both authors working in the Research Division of the Minnesota State Board of Control. The thirty-nine tests comprising the battery are arranged in a scale of overlapping units, the net results of which closely approximate the results from any good individual mental examination. The tests in their most recently revised form are arranged in the booklets as shown in Table 22.

TABLE 22

Arrangement of Kuhlmann-Anderson Tests

School Grade		Tests	Age When Test Fits Best
I	First semester	1–10	6–0
I	Second semester	4–13	6–6
II		8–17	7–6
III		12–21	8–6
IV		15–24	9–6
V		18–27	10–6
VI		22–31	11–6
VII-VIII		25–34	13
IX-XII	and adult	30–39	15–6

[3] Kuhlmann, F., and Anderson, Rose, *Kuhlmann-Anderson Intelligence Tests*, Educational Test Bureau, Minneapolis, Minnesota, 1927.

A somewhat novel procedure is used in interpreting the results of these tests. Each of the ten tests comprising a booklet is standardized separately. The test is scored in terms of the number of exercises answered correctly. By referring to a table of norms the mental age of the individuals making such a score is obtained. A mental-age score for each of the ten tests is thus obtained, the final mental-age score assigned to the pupil being the median of the resulting mental ages. This procedure appears to result in unusually reliable measurement of mental ability.[4]

2. Otis Group Intelligence Scale [5]

Advanced Examination, Forms A, B

This was one of the first tests to appear for measuring intelligence at the secondary-school level. It has been widely used in Grades VII to XII inclusive. It is composed of ten divisions as follows: following directions, opposites, disarranged sentences, proverbs, arithmetic, geometric figures, analogies, similarities, narrative completion, and memory. The test requires more than an hour to give; it has 230 test elements with an actual working time of about 45 minutes. The coefficient of reliability for grades and half grades is around .84 and around .97 for all grades combined. The test correlates approximately .75 with a suitable criterion.

3. Self-Administering Test of Mental Ability [6]

Higher Examination, Forms A, B, C

This test is unique in that it requires a minimum amount of instruction from the examiner. For this reason industrial education teachers will find this a very satisfactory test to use. The test has two time limits of 20 and 30 minutes. Generally the 30-minute time is satisfactory except possibly in the last years of the senior high school. The reliability of the test is reported as .92, and it has a high correlation with a valid criterion.

[4] Kuhlmann, F., "A Median Mental Age Method of Weighting and Scaling Mental Tests," *Journal of Applied Psychology,* June, 1927.

Pintner and Patterson, *A Scale of Performance Tests,* 1917.

[5] Otis, A. S., *Group Intelligence Scale,* World Book Company, Yonkers-on-Hudson, New York, 1918.

[6] Otis, A. S., *Self-Administering Test of Mental Ability,* World Book Company, Yonkers-on-Hudson, New York, 1922.

4. Terman's Group Test of Mental Ability [7]

Forms A and B

The Terman test is composed of ten divisions as follows: information, best-answer, word meaning, logical selection, arithmetic, sentence meaning, analogies, mixed sentences, classification, and number series. Two approximately equal and interchangeable forms are available. There are 185 items in each form of the test. It can be given in 40 minutes, although the actual working time is only 27 minutes. The reliability of the test is approximately .90. It correlates .75 with a suitable criterion of mentality. Complete tables of mental-age norms are given in the examiner's manual.

45. Individual Mental Examinations.

Individual mental examinations probably constitute the most accurate devices for the measurement of intelligence. The length of the test, the wide variety of reactions called for, the fact that the subject receives his instructions personally from the examiner, and the fact that this affords the examiner an opportunity to observe each reaction of the subject all combine to account for this greater accuracy of measurement. However, this greater accuracy is compensated for by greater expense in the administration of the tests, which operates in terms of both time and money. In fact, it is thought by many that this expense item is so great that in most classroom and shop situations the resulting increase in accuracy of measurement is not commensurate. Accordingly, it is quite probable that teachers of industrial education will find it desirable to initiate their analysis of problem cases by first using good group tests of mental ability. Individual mental examinations may be given later to a relatively small number of pupils who deviate most widely from the normal.

A very simple procedure will reveal directly to the teacher the special individuals who should receive further attention. If the group mental-test scores for the entire class are tabulated in a frequency table, or if the test papers themselves are merely arranged in descending order of size of the test scores, the individual pupils deviating most widely from the average for the class and from the normal mental age for the grade will be revealed. Thus, it may be necessary to retest (or to give the individual mental examination to) only a small percentage of the group. After all, it is in the highest and lowest ex-

[7] Terman, L. M., *Group Test of Mental Ability,* World Book Company, Yonkers, New York, 1920.

tremes of intelligence that the problem cases arise, and it is also in this same group that the most serious errors or misinterpretations are likely to take place. Most present-day tests of mental ability accomplish a reasonably accurate placement of the more nearly normal group.

STANFORD REVISION OF BINET-SIMON INTELLIGENCE TESTS [8]

This extensive mental scale includes groups of tests suitable for the measurement of mental ability from three years to fourteen inclusive as well as tests suitable for average adults and superior adults. There is a complete manual of directions, stating in detail just how to administer and score the test. The reliability of the test is around .95, and its validity is commonly considered as a standard in constructing other group and individual tests of intelligence. The validity of the test has been frequently criticized because it is composed of so much verbal material, but this same criticism can be applied to many intelligence tests. This is one of the most widely used of the individual tests of intelligence for classroom use.

46. Results of Mental Measurement.

Intelligence-test scores should be interpreted with great care. After all, such scores are only estimates of intelligence and should not be considered absolute measures. The individual is very complex, and many factors may affect the rating of the pupil. In the first place, the differences in environment and in training opportunities are frequently overlooked in interpreting mental-test scores. The intelligence of different individuals may legitimately be compared only when there is assurance that the learning opportunities have been the same. This fact, incidentally, is difficult to establish. Accordingly, few such comparisons are legitimate. Numerous other factors such as the individual's inability to see, to read, to hear, or some other temporary physical disability may seriously influence the score. Errors in the administration of the tests, errors in scoring, and clerical errors in transcribing and in computing results must be guarded against at all times. The shop teacher should be especially critical of very high and very low scores since they are the ones most likely to be in error. Every very low and every very high score should be carefully

[8] Terman, L. M., *et al.*, *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence,* Warwick and York, Baltimore, 1917. Test material also through Houghton Mifflin Company, Boston, and C. H. Stoelting Company, Chicago.

rechecked by other group tests or by individual examinations before any serious administrative or instructional adjustments are made.

47. Mental-Age Score.

The results of mental testing which are of most use to the classroom teacher are the mental-age scores. These scores are derived from the raw test scores and afford the basis for the calculation of a number of useful derived scores called quotients. Mental-age norms for an intelligence test are commonly secured by administering the mental examination to large numbers of individuals of various age levels. After the tests have been scored the papers are usually assembled by age groups, all the nine-year-olds being placed in one group, all the ten-year-olds in another, etc. In this way the typical scores to be expected of individuals of different chronological ages may be determined. If the typical point score of a large group of ten-year-olds should be 126 on a given mental test, thereafter any individual making a score of 126 points would be designated as having a mental age of ten years. Certain of the more carefully validated and standardized group tests have established their mental-age norms on the basis of results from large numbers of individual mental examinations.

48. Intelligence Quotients.

Mental-age scores make possible the derivation of a series of quotients which are very useful in the interpretation of mental- and achievement-test results. The most commonly used quotient of this type is the intelligence quotient or I.Q. The I.Q. is the ratio of the mental age to the chronological age of the individual tested. The formula for the I.Q. is $\text{I.Q.} = \frac{\text{M.A.}}{\text{C.A.}}$, in which the M.A. is the mental age and the C. A. is the chronological age, both expressed in months. The I.Q. itself expresses the relative mental development of the individual. If the pupil makes a score on the mental test which gives him a mental age identical with his life age his resulting I.Q. is 1.00, or 100 as it is usually expressed. A pupil is commonly considered normal if his I.Q. falls between 90 and 110. Intelligence quotients of 110 to 120 are above average, and above 120 are superior. Quotients above 130 approach the genius class, and quotients of 140 to 150 indicate most unusually accelerated mental development. Similarly, quotients of 80 to 90 are below average. Individuals with quotients of less than 80 are poor and may be expected to encounter much difficulty in the mastery of abstract material at the junior- and senior-high-school level. Quotients of 70 to 80 are very low, and quotients

pelow 60 indicate exceedingly retarded mental development bordering on the moron level and idiocy.

In the interpretation of the quotients derived from group-mental est scores it should be remembered that such quotients represent individual interpretations, whereas, as a matter of fact, the tests on which they are based are group tests. In general, intelligence quotients based on group-test results should not be utilized for any serious purpose on an individual basis.

Intelligence-test scores are generally regarded as professional information to be used in teaching and guidance, but not to be given to the pupil or his parents. Long experience with intelligence tests has proved this to be a wise policy. The scores from such tests are suggestive to the teacher and should be used only as indicative of capacity. When these ratings are given to the layman he is likely to look upon them as final rather than suggestive and fail to interpret them in the light of a professional background. There may be rare occasions when it is feasible to give this information to a pupil or his parents provided it will aid the pupil or parent in a better understanding of the pupil's possibilities of accomplishment or his future needs. It is possible sometimes to encourage a pupil who is doing poor work and is discouraged by pointing out to him that he has average native ability and can succeed with proper application. Occasionally it may be well to point out to a dull pupil that he is doing well in the light of his ability. It may be that a lazy, bright pupil can be motivated by pointing out his failure to capitalize his real ability. Occasionally, parents who punish their children for low marks can be made a little more lenient by showing them that their children are doing well for their ability. These suggestions must be used with extreme tact and care or they will prove destructive rather than constructive.

The shop teacher must be careful to distinguish clearly between intelligence-test scores and tests of achievement. The intelligence test gives an indication of the student's capacity for acquiring information largely through the use of abstract processes. The achievement test aids in the measurement of actual accomplishment in the class and can be used as a partial basis for assigning shop marks. The intelligence-test score is of value in teaching and guidance; it should not be used directly in marking achievement inasmuch as intelligence tests are not measures of achievement in specific courses.

II. MEASUREMENT OF SPECIAL APTITUDES

49. Measurement of Special Abilities.

The measurement of general mental ability suggests the possibility of securing objective evidence of special types of abilities or aptitudes. This is a field of measurement in which all teachers in the secondary schools should be interested, for adequate educational guidance is becoming more and more important at these levels in our educational program. Educational and vocational misfits, the high pupil mortality in certain of our courses, the heavy teacher-burden caused by increasingly large classes, as well as the general embarrassment to the school resulting from the misapplication of abilities, all demand that more attention be given to this phase of educational measurement. Industrial education teachers, because of their proximity to the problems, represent a group which should be greatly interested.

50. Measuring Mechanical Aptitude.

An aptitude may be thought of as the capacity of an individual for the development of some special ability or skill. Mechanical aptitude is the capacity of an individual to deal successfully with mechanical devices, and to acquire the knowledge essential to their selection and operation after suitable training has been given. An individual who has a large measure of mechanical aptitude, other things being equal, will readily succeed when given instruction. On the other hand, an individual with low mechanical ability is likely to fail regardless of the instruction or opportunities given to work with mechanical things.

The importance of identifying mechanical aptitudes is more obvious when it is realized that at least 40 per cent of the gainfully employed population in the United States is dependent in some measure for its economic success on the possession of mechanical ability. It is true, of course, that mechanical ability is only one factor in success even in mechanical pursuits, but it is also true that it is quite an important factor. The industrial education teacher should keep this clearly in mind and should blend other important guidance information with the evidence of the pupil's mechanical ability.

It thus becomes apparent that a knowledge of the student's mechanical ability is important to industrial education teachers from both the guidance and the instructional points of view. Knowledge of the fact that an individual has low or high mechanical aptitude gives the industrial education teacher an objective basis for guiding the pupils into or out of vocations which involve high degrees of these abilities. It enables the teacher to assign projects better adapted to

the individual differences of the pupils in the class. Such a knowledge is of value to trade- or continuation-school teachers in selecting individuals who are likely to profit by the instruction offered. However, it is well to bear in mind that mechanical-ability tests must be carefully administered and interpreted, and that, at best, they are merely very suggestive and should not be considered infallible.

It is widely known that individuals vary in mechanical ability. It is also known that mechanical ability does not correlate highly with intelligence, the quotient usually being around +.40. Stenquist pointed this out a number of years ago. This does not mean that many individuals with high intelligence as measured by intelligence tests do not have high mechanical ability, nor does it mean that individuals with low intelligence always have high mechanical ability. It strongly suggests that there may readily be a concrete aspect of intelligence which is necessarily an accompaniment of intelligence of the abstract type. Paterson, Elliott, *et al.*[9] report that there is a fairly uniform tendency for test scores on mechanical ability to increase with chronological age between eleven and twenty. The same authors found little support for the supposition that men excel women in mechanical ability. The only test in which men and boys clearly excelled was in the *Minnesota Assembly Test,* and this was probably due to greater experience with the materials. Judging from the data available, engineering students are not superior to liberal arts students in innate mechanical ability. This emphasizes the fact that guidance is made up of many important factors and even in engineering colleges mechanical ability is not an infallible guidance factor.

Industrial education teachers can readily see the guidance value of tests of mechanical ability, and fortunately some good tests are available for use. The three measures of mechanical ability described and discussed in the following pages deserve careful study by shop and drawing teachers.

1. Minnesota Mechanical-Ability Tests[10]

The *Minnesota Mechanical Ability Tests* are the outcome of research at the University of Minnesota during the years 1923-1927. They are probably the most carefully prepared tests of mechanical aptitude that have been published for general use. The tests are quite

[9] Paterson, Elliott, Anderson, Toops, *Minnesota Mechanical Ability Tests,* University of Minnesota Press, Minneapolis, Minnesota, 1930, pp. 282-284.

[10] Materials for the *Minnesota Mechanical Ability Tests* may be obtained from the Marietta Apparatus Company, Psychological Equipment, Marietta, Ohio.

reliable, and their validity has been carefully checked against objective criteria.

When administered according to directions the *Minnesota Mechanical Ability Tests* will give results which will be very useful in teaching and guidance. The battery includes the following six tests:

(*a*) Minnesota Paper Form Board, Series A and B, for either group or individual testing.

(*b*) Minnesota Spatial Relations Test (individual test).
  Boards A and B.
  Boards C and D.

(*c*) Minnesota Assembly Test (group or individual).
  Long form.
    Box A, B, and C.
  Short form.
    Box 1 and 2.

(*d*) Minnesota Interest Analysis Test.

(*e*) Packing Blocks Test.

(*f*) Card Sorting Test.

The authors report the following coefficients of reliability and validity for these tests.

TABLE 23

COEFFICIENTS OF RELIABILITY AND VALIDITY ON MINNESOTA MECHANICAL ABILITY TESTS [11]

Test	r_n*	Validity Coefficient (Between test and quality criterion)
Minnesota Assembly, Boxes A, B, C........	.94	.55
Minnesota Paper Form Board, A and B.....	.90	.52
Minnesota Spatial Relations, Boards A, B, C, D	.84	.53
Packing Blocks	.77	.26
Card Sorting	.90	.19

* $r_{\frac{1}{2}\frac{1}{2}}$ stepped up by Spearman-Brown formula.

The manual of directions gives instructions for administering, scoring, and interpreting results. The authors state that the examiner need not be a trained psychologist to administer the tests, but that he should be thoroughly familiar with the test to give it successfully.

[11] *Op. cit.*

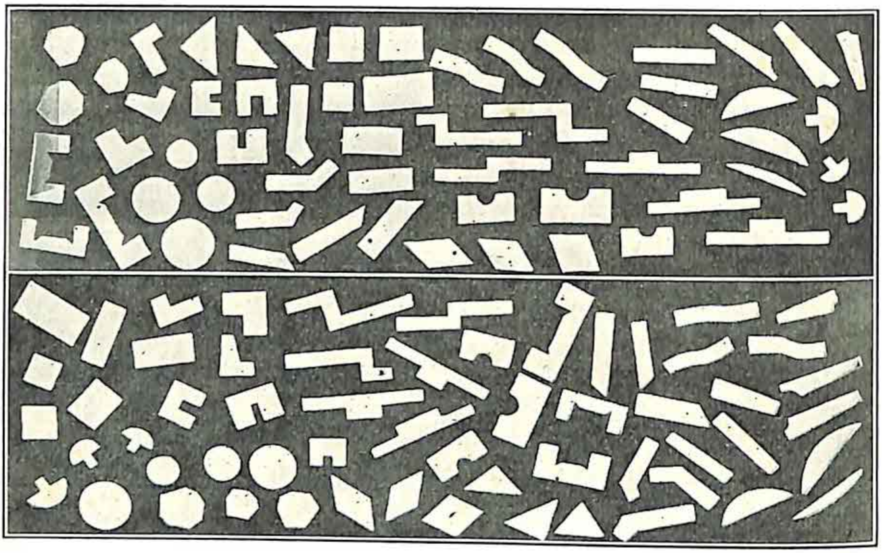


FIG. 4.—Minnesota Spatial Relations Test.

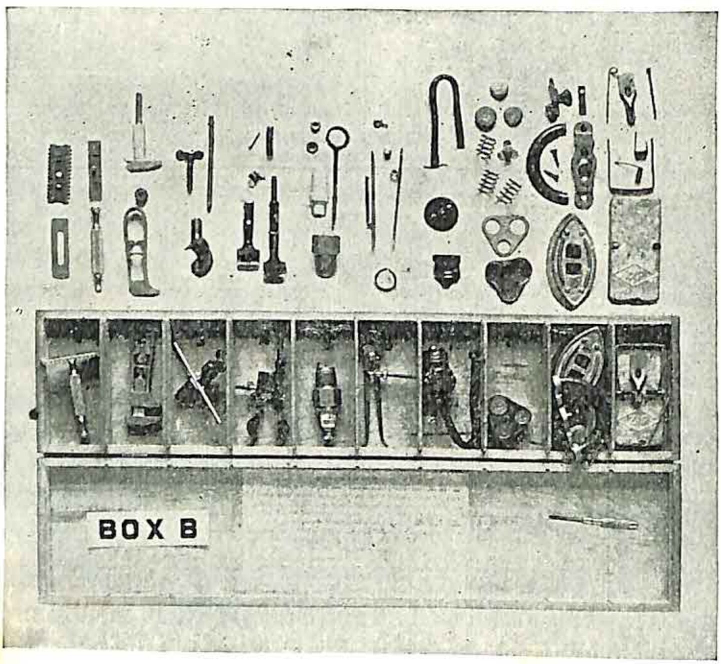


FIG. 5.—Minnesota Assembly Box Test.

The test is quite elaborate and rather expensive in time and money, but the combined battery will yield a satisfactorily valid and reliable score for most guidance purposes.

2. Stenquist's Assembling Tests of General Mechanical Ability [12]

This is one of the first assembling tests of mechanical ability to be widely used. The test is composed of a small rectangular box divided into ten compartments. Each compartment contains a small mechanical device which is common in the experience of most people. Some of the items selected by the author are a mouse trap, push button, simple lock, and bicycle bell. These mechanical devices are arranged in order of difficulty of response. In scoring, ten points are allowed for each device, and the score on each item is the number of correct points in assembling the device. Stenquist also suggests a short method in which just the devices almost or completely assembled are counted and given as the score. The second method would be less reliable because it disregards partial accomplishment. The test also gives a bonus of one-half point per minute for each minute under 30 minutes in responding to the test.

According to the literature this test has a reliability of .70 and correlates with intelligence-test scores about +.20 to +.30. The test measures certain aspects of mechanical ability, but has too low a reliability to be used with assurance in measuring mechanical ability in individual cases. It correlates with teachers' marks in mechanical subjects as high as +.80. Paterson, Elliott, and others found that the test correlated +.26 with the objective criterion used in the validation of the *Minnesota Tests of Mechanical Ability*. They also found that, by increasing the length of the test, its reliability and validity were greatly improved. Norms for different age groups are given in the accompanying table.

TABLE 24

Percentile Norms

Age	Percentiles 5	10	25	50	75	90	95
Fifteen	1.0	1.7	3.2	4.6	6.2	7.2	7.9
Fourteen	1.0	1.4	2.4	4.3	6.0	7.4	8.0
Thirteen	1.0	1.5	2.5	3.9	5.3	6.6	7.7
Twelve	.7	1.0	1.8	2.9	3.8	5.2	6.8

[12] C. H. Stoelting Company, Chicago.

3. Stenquist Mechanical Aptitude Tests

Tests I and II

These pencil-and-paper tests of mechanical aptitude are made up of pictures of mechanical things which are common in the experience of most people. The tests are not two equivalent forms, but the two parts (I and II) are to be used to supplement each other. In general the student taking the test has to recognize mechanical things that belong together or work together and answer questions about parts or operations of machines. The working time on the first test is 45 minutes, and on the second 50 minutes. The two forms together require 173 responses.

The reliability of the test appears to be around .75. Paterson and Elliott have shown that this can be increased to .89 or .90 by increasing the length of the test. The validity as checked by the best known criterion is lower than the assembly test even when the reliability has been corrected. Correlation with the objective criterion of the Minnesota tests is around +.30. The test correlates as high as +.60 with teachers' marks in shop courses.

III. SUMMARY

The methods of measuring general and special types of mental abilities are discussed in this chapter. Intelligence, as treated in this discussion, is considered to be the capacity for learning, plus the information, skills, and attitudes which the individual has gained from reacting to his environment. Certain individuals react readily to abstract stimuli; others respond most readily to concrete and tangible situations. For this reason there seems to be a real need for both abstract and concrete types of mental stimuli.

Intelligence tests are commonly classified as individual mental examinations and group tests of mental ability. The results of mental testing which are of most use to the classroom teacher are the mental-age scores. These and all other scores derived from mental tests should be regarded as professional information of the most confidential type and used accordingly.

An aptitude is the capacity of an individual to develop special abilities or skills. Mechanical aptitude represents the potential ability of the individual to deal successfully with mechanical devices, and the knowledge essential to the selection and operation of such devices after a suitable period of training. An early knowledge of special

aptitudes on the part of individual pupils is of great importance to the teacher of industrial education courses.

SUMMARY EXERCISES FOR DISCUSSION

1. State what seems to you to be a practical and accurate definition of intelligence.
2. List the outstanding advantages and disadvantages of group mental tests.
3. What are the major advantages and disadvantges of the individual mental examination over the group test?
4. What is the difference between the mental age and the I.Q.?
5. Show how the shop teacher needs mental-test results as a protection against the possible misinterpretation of achievement-test results.
6. What is the basis for the statement that aptitude tests are tests of special types of intelligence? Is it true?
7. Why should the shop teacher be especially concerned with results from aptitude tests?
8. From available sources make a list of the mental tests and special-aptitude tests which would seem to provide the most useful information for the industrial education teacher.

SELECTED REFERENCES

BIRD, VERNE A., and PECKSTEIN, L. A., "General Intelligence, Machine Shop Work, and Educational Guidance in the Junior High School," *School Review*, Vol. 29: 782-6.

BAKER, HARRY J., and CROCKETT, A. C., *Detroit Mechanical-Aptitude Examination*. Bloomington, Illinois: Public School Publishing Company, 1928.

BOARD, EDNA; MARSH, WILLA; and STOCKWELL, LYNN E., "Relation of General Intelligence to Mechanical Ability," *Industrial Arts Magazine*, Vol. 16: 330-2, September, 1927.

CARPENTER, J. E., "The Function of Mental Tests in the Administration and Supervision of a Vocational School," *Vocational Education Magazine*, Vol. 2: 65-6, September, 1923.

FREEMAN, F. N., *Mental Tests, Their History, Principles and Applications*. Boston: Houghton Mifflin Company, 1926.

FRYKLUND, VERNE C., "Intelligence and the Shop," *Industrial Arts Magazine*, Vol. 17: 81-3, March, 1928.

GORDON, GEO., JR., "Relation of Pupils' Intelligence Quotients to Their Grades in High School Shops," *Industrial Education Magazine*, Vol. 30: 249-50, January, 1928.

HENIG, M. S., "Intelligence and Shop Accidents," *Industrial Arts Magazine*, Vol. 18: 265-6, August, 1928.

HERRING, J. P., *Herring Revision of the Binet-Simon Tests*. Yonkers, New York: World Book Company, 1922.

HULL, C. L., *Aptitude Testing*. Yonkers, New York: World Book Company, 1928.

KEANE, FRANCIS L., and O'CONNER, JOHNSON, "A Measure of Mechanical Aptitude," *Personnel Journal*, Vol. 6: 15-24, January, 1927.

KUHLMANN, F., and ANDERSON, ROSE, *Examiner's Manual for Kuhlmann-Anderson Intelligence Tests.* Minneapolis: Educational Test Bureau, Inc., 1927.

MACQUARRIE, T. W., *A Test for Mechanical Ability.* Hollywood: Southern California School Book Depository, 1925.

MADSEN, I. N., "The Contributions of Intelligence Tests to Educational Guidance in High School," *School Review,* Vol. 30: 692-701, November, 1922.

MEIER, N. C., and SEASHORE, C. E., *Art-Judgment Test.* Iowa City, Iowa: Bureau of Educational Research, University of Iowa, 1929.

Minnesota Paper Form-Board Test, Series A and B. Marietta, Ohio: Marietta Apparatus Company, 1920.

OTIS, ARTHUR S., *Group Intelligence Scale, Examiner's Manual.* Yonkers, New York: World Book Company, 1918.

O'ROURKE, L. J., *Mechanical Aptitude Test.* Washington, D. C.: The Educational and Personnel Publishing Company, 1926.

PROCTOR, W. M., "Psychological Tests and Guidance of High School Pupils," *Journal of Educational Research Monographs,* No. 1. Bloomington, Illinois: Public School Publishing Company, 1923.

REEDY, CAROLINE M., "Can Intelligence Tests Help to Solve the Continuation School Classification Problem?" *Vocational Education Magazine,* Vol. 2: 151-5, October, 1923.

RUTH, NORTON W., *Electrical-Inclination Test.* Chicago: C. H. Stoelting Company, 1927, 16 pages.

SEASHORE, C. E., and others, "Mentality Tests: a Symposium," *Journal of Educational Psychology,* Vol. 7: 229-40; 278-86; 348-60; April, May, June, 1916.

SMITH, H. L., and WRIGHT, W. W., *Tests and Measurements,* Chapter XX, "Prognosis and Special Ability Tests." New York: Silver Burdett and Company, 1928.

STENQUIST, JOHN L., *Assembling Tests of General Mechanical Ability,* Series I, II, and III. Chicago: C. H. Stoelting Company, 1922.

STENQUIST, JOHN L., *Mechanical Aptitude Tests.* Yonkers, New York: World Book Company, 1927.

STOY, E. G., "Tests for Mechanical Drawing Aptitude," *Personnel Journal,* Vol. 6: 93-101, and 361-6, July, 1927, October, 1927.

SUTHERLAND, S. S., "Correlation Between Intelligence and Skill in Shop Work," *Industrial Arts Magazine,* Vol. 17: 204-5, June, 1928.

TERMAN, L. M., *The Intelligence of School Children.* Boston: Houghton Mifflin Company, 1919.

TERMAN, L. M., and others, *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence.* Baltimore, Maryland: Warwick and York, 1917.

CHAPTER IX

TESTS IN RELATED EDUCATIONAL FIELDS

51. Relation of Other Educational Achievement to Industrial Arts.

Achievement in industrial education cannot be measured adequately without the supplementary information procurable only through the use of educational tests selected from other, related fields. Just as it is impossible to give a proper interpretation to the results of achievement testing in any field of subject-matter without the use of such supplementary information as mental tests provide, so it is difficult if not impossible to secure a complete evaluation of instruction in the industrial arts without the use of definite and accurate measures of the many factors which contribute to achievement in these subjects.

Achievement in the content subjects is limited to a very high degree by the student's reading ability. The comprehension of the precise types of directions, symbols, and instructions given in the industrial arts subjects is basic. Certain skills in arithmetic, algebra, and the sciences are essential in shop work. Mastery of certain English usages and mechanics is as essential to acceptable achievement in this field as in almost any other. A reasonable skill in freehand drawing, lettering, and handwriting is also an important limiting factor in industrial arts achievement. In this chapter a number of the more useful educational tests selected from important fields related to industrial education are described.

52. Reading Tests.

The development of the ability to read is one of the most important educational and vocational accomplishments of the school. Achievement and school progress depend to a very large degree upon reading ability, and the higher up in the grades the pupil progresses the more important does reading ability become. In fact, in the high school and college, reading ability is the most important single means by which knowledge and information are secured. The recognition of silent reading as a basic study tool has done much to improve the quality of initial instruction given in the subject. It has also resulted in stimulating the development and use of effective remedial instruction in this field.

The elementary school is expected to develop the general reading skills, including an effective comprehension of the content of material read at an economical and efficient rate. Unduly slow reading is a handicap, just as is poor comprehension. However, the elementary school deals with reading problems involving comparatively common usages and simple vocabularies. It does not concern itself especially with technical terms and symbols used in the special subjects, such as the industrial education courses. Thus it becomes largely the responsibility of industrial education teachers to train their pupils to read the technical phases of their specific subjects. One of the most effective ways to accomplish this is for the teacher to determine at once the general reading ability of his class. In any event, it is quite certain that the pupil must have a general reading ability before he can acquire the technical reading ability required in many of the specialized industrial education courses.

A number of excellent general tests of silent reading ability are available, but thus far no one has developed a particularly outstanding device for the measurement of the pupil's ability to read the technical content of specialized courses. Special reading tests designed to parallel the teaching in the industrial education subjects are badly needed. Reading tests using technical content selected from the fields of woodworking, drawing, sheet-metal working, auto mechanics, electricity, printing, etc., would be of great value.

The two reading tests described in this chapter are primarily tests of reading ability as it exists at the secondary-school level. The tests have been selected as being best suited to meet the needs of teachers and supervisors of industrial education courses in securing objective and detailed analytical information concerning the different aspects of the reading ability of their pupils. The results from the use of the tests and the general specifications upon which they are built should prove helpful to industrial arts teachers in constructing and using technical reading tests for the various industrial education subjects.

1. Haggerty Reading Examination, Sigma 3
Forms A and B [1]

This distinctly valuable test is designed to measure general silent reading ability from the fifth grade through the twelfth. The total score on the test indicates a measure of general reading comprehension. The test is composed of three subtests of vocabulary, sentence reading,

[1] Haggerty, M. E. and Laura C. *Reading Examination, Sigma 3,* World Book Company, Yonkers, New York, 1920.

and paragraph reading. The vocabulary content of the test was validated by selecting the common words used in seventh- and eighth-grade readers and history texts. The actual working time is 28 minutes, but the administration of the test requires about 45 minutes.

2. Iowa Silent Reading Test; Advanced Forms A and B (Revised) [2]

This test is designed to secure an analytical measurement of the silent reading skills used in reading of the work-study type. By the use of a series of tests sampling into several different types of reading skills the total comprehension score is intended to reveal general reading ability. The scores on the separate test parts afford the basis for the analysis of the strengths and weaknesses of individual students. The several parts of the test cover the following unit skills which contribute to the student's ability to read and to work with books:

Test 1. Paragraph meaning.
- Science material.
- Literary material.

Test 2. Word meaning.
- Social science vocabulary.
- Science vocabulary.
- Mathematics vocabulary.
- English vocabulary.

Test 3. Paragraph organization.
- Selection of central idea.
- Outlining.

Test 4. Sentence meaning.

Test 5. Location of information.
- Use of the index.
- Selection of key words.

Test 6. Rate of silent reading.

The total working time for this test in its recently revised form is 35 minutes. This makes it easy to administer the test within a class period. In spite of this relatively short working time the reliability of the test is high. The reliability coefficients and the P.E.'s of scores reported in Table 25 indicate that scores on these tests may be taken as very accurate measures of the silent reading abilities of high-school or college students.

[2] Greene, H. A., Jorgensen, A. N., and Kelley, V. H., *Iowa Silent Reading Test, Advanced.* Revised Edition for High Schools and Colleges. World Book Company, Yonkers, New York, 1931.

TABLE 25

RELIABILITY OF IOWA SILENT READING TEST, ADVANCED *

Test	Lowest			Highest		
	r	P.E. Score	Grade	r	P.E. Score	Grade
1	.84	.65	9	.91	.62	13
2	.87	1.68	9	.94	1.08	11
3	.86	.53	11	.94	.64	12
4	.59	.94	9	.95	.65	10
5	.80	.94	9	.90	.63	13
Total Compr.	.95	5.73	12	.96	4.98	9

* Quoted from table in examiner's manual.

53. English Tests.

The social importance of the use of correct language habits is so great that no teacher can afford to relax for a moment in his demands for correctness in the oral and written language of his students. Teachers of industrial education must share this responsibility in spite of the facts that very often this is a field outside their specific realm of interest, and that the written language used in their courses is usually quite limited. The demands made on the oral language skills are usually as extensive, however, as they are in other subjects. The professionally minded teacher of the industrial education subjects will be just as careful to watch and correct the language habits of his pupils as he is to watch his own language usages. Correct habits of speech and writing come only through extensive and continuous practice in correct usage. The industrial arts teacher must be in a position to cooperate at all times with the English teachers, whose major responsibility it is to see that these correct habits function in school and the situations of daily life.

Achievement in English expresses itself in a number of different ways, most of which are measurable with tests of some type. The most commonly measured phases of English are in the fields of word usage, grammar, and the mechanics of written composition. General merit of the total written language production is also measurable with the help of certain quality scales. Thus far, oral composition has evaded most attempts to objectify its measurement. A few tests and scales selected from the large number available in this subject are described and evaluated here from the standpoint of their use to the teacher of the industrial arts subjects.

Measurement of Composition. English composition is ordinarily measured in one of two ways. One method concerns itself with the more or less objective evaluation of the general qualities of merit which the production possesses. This procedure makes use of a scale composed of written productions of different degrees of merit which have been scaled or evaluated and arranged in ascending order of merit in accordance with a numerical scale. Each different specimen is assigned a numerical value in terms of the relative merit it possesses. In general, the lower the merit of the specimen the lower the quality scores assigned to it. In actual use the scale is not taken into the classroom, but is used by the teacher as a means of assigning general merit ratings to the written productions prepared by students under rather carefully controlled conditions, and on selected topics and subject-matter. The other method of measuring composition is by checking its form and freedom from mechanical errors. One of the very useful scales for the measurement of English composition, the *Willing Composition Scale,*[3] makes use of both these procedures.

The compositions, after being rated for story value or general merit, are rated for form value by making a careful check of all spelling, punctuation, capitalization, grammatical, and usage errors. The total number of such errors is then divided by the number of words comprising the composition. This result is then multiplied by 100, which expresses these form errors in terms of the number of such errors per 100 words of composition. The number of form errors per 100 words declines as the quality or story value of the composition rises.

The *Willing Scale* is the only one of the commonly used scales which makes any attempt to combine the ratings for form and quality. The *Thorndike Extension of the Hillegas Scale,* the *Hudelson Composition Scales,* the *Nassau County Extension (Trabue) of the Hillegas Scale* are all very useful general merit scales but confine their measurement entirely to composition merit. It seems quite likely that most industrial arts teachers interested in making any intensive check on the merit of the written work of their students will find the form values and the story values resulting from the use of the *Willing Scale* the most useful measures.

Measurement of Grammar. Two somewhat contrasting types of grammar tests are described in this section.

[3] Willing, Matthew H., *The Willing Composition Scale,* Public School Publishing Company, Bloomington, Illinois, 1918.

The *Iowa Grammar Information Test* [4] is resigned to meet the need for a test of the purely informational aspects of English grammar. In addition to its survey use in English classes, it should prove to be a valuable measure of the formal background of grammar needed by students who are beginning the study of a foreign language. By means of 80 objective exercises of the three-answer multiple-choice type it samples into almost all the commonly taught phases of English grammar. Two equal and parallel forms are available. Percentile norms are based on 1557 cases in Grades VII to XII.

The *Kirby Grammar Test* [5] is intended to be used in the measurement of usage and grammatical errors in Grades VII to XII. The pupil is tested on his knowledge of verbs, pronouns, and certain miscellaneous usages. For convenience in administration, the exercises are arranged in five divisions each containing usage exercises of the alternate-response type. The pupil is required to select the correct form for a given exercise and then to indicate (by recognition) the grammatical rule which governs its use.

The reliability of the score on the principles test is about .90, but on the sentence test is only around .60. Norms are given for Grades VII to XII, but there is not a great difference between the norms for the different grades. This seems to indicate that pupils do not improve much in grammar during their secondary-school work. The actual working time of the test is about 35 minutes.

Language Usage. The language-usage tests described in this section illustrate two different types of measurement in language. The first is an analytical test sampling many different language abilities. The second is a general survey of language usage based on the recognition of error.

The *Iowa Elementary Language Tests* [6] are designed for survey purposes in Grades IV to IX inclusive. However, the reliability of measurement on the different parts permits a very useful type of analysis of language limitations. The eleven phases of language ability sampled by this test range over a total of 304 different items with a total possible score of 338 points. In Test 1 which deals with two phases of word meaning the four-answer multiple-choice type of

[4] Cram, Fred D., and Greene, H. A., *The Iowa Grammar Information Test,* Bureau of Educational Research and Service, Extension Division, University of Iowa, Iowa City, 1935.

[5] Kirby, Thomas J., *The Kirby Grammar Tests,* Bureau of Educational Research and Service, Extension Division, University of Iowa, Iowa City, Iowa, 1920.

[6] Greene, H. A., Ballenger, H. L., *Iowa Elementary Language Tests,* Educational Test Bureau, Minneapolis-Philadelphia, 1929.

exercise is used. Alternate-response exercises are used in two of the three tests measuring phases of language usage. The recognition-correction type is used in Tests 2-B, 6-A, and 6-B. A novel type of technique utilizing keyed brackets is used in Test 5.

The *Wilson Language Error Test*[7] is available in two parts, each consisting of three forms. The forms consist of short stories of about 300 words which contain a number of common language errors. The pupil is to read the story and correct the language errors. The test is simple to administer and, when at least three forms are used, has valuable diagnostic power. The errors included in the tests are those commonly made as indicated by studies of pupils' errors in several different schools. The reliability of the test is about .80. The test should prove valuable to industrial education teachers in diagnosing common language errors. Norms are given for Grades VII to XII but they show approximately the same levels of achievement in all the grades.

54. Vocabulary Tests.

Several good general vocabulary tests have been developed. These are of some value to teachers of industrial education but they are included here more for the suggestions they give concerning methods that may be employed in developing suitable vocabulary or word meaning tests to parallel the different industrial education courses. *Hunter's*[8] *W-4 Trade Names Test in Woodwork* is one of the pioneer efforts of industrial education teachers to develop tests along the line of technical vocabulary. The inability of a pupil to understand the meaning of words used in a given course does not necessarily mean that he should not take the course, but indicates the need for special instruction and drill in word meaning early in the course so that the pupil can better profit from the instruction.

The *Pressey Technical Vocabularies of the Public School Subjects*[9] should be most suggestive to industrial education of possible methods of developing tests upon the technical vocabulary used in the industrial education subjects. This vocabulary list, which includes technical vocabularies for fifteen school subjects, contains a list of technical words pertaining to woodwork and elementary metal work, but it is not entirely adequate for industrial education purposes because it covers only these two courses.

[7] Wilson, G. M., *Wilson Language Error Tests*, World Book Company, Yonkers, New York, 1923.

[8] Hunter, W. L., *Shop Tests*, The Manual Arts Press, Peoria, Illinois.

[9] Pressey, Luella C., *Technical Vocabularies of the Public School Subjects*, Public School Publishing Company, Bloomington, Illinois, 1923.

The procedure in selecting and rating the technical vocabularies was as follows. First, all unusual or technical words which appeared in commonly used textbooks of the subjects treated were tabulated and classified; second, the terms were rated according to importance by a group of special teachers of the subjects; and third, terms were classified as essential that were checked by more than half of the teachers. This is in no sense a test but is a suggestive vocabulary study for further development in the industrial education field.

55. Spelling.

Industrial education teachers have a distinct responsibility to teach the pupils in their classes to spell the technical words peculiar to their courses and to aid the other teachers in the school in maintaining proper spelling levels in written work. To equip the child with a method of learning to spell and to teach the spelling of commonly used words is the specific function of the elementary school, but it takes continual cooperation by teachers of all subjects to assure the lasting assimilation and mastery of these fundamental skills. Teachers of industrial education should recognize this responsibility.

The majority of available spelling tests and scales have been developed for the elementary school. However, at least two such spelling scales are definitely designed for use at the secondary-school level. These two scales which are briefly described here should prove of very definite value to industrial education teachers in measuring general spelling ability on the junior- and senior-high-school levels. They should also prove suggestive for the construction of spelling scales dealing with the technical words which are an integral part of instruction in industrial education.

1. Sixteen Spelling Scales Standardized in Sentences for Secondary Schools [10]

These scales, frequently called the *Seven-S Scales*, consist of 16 separate and scaled lists of 20 words each. It requires about 5 minutes to give any one of the scales, and for individual scores it is advisable to use two and combine the scores. The tests have been carefully prepared and afford a very satisfactory means for industrial education teachers to measure the general spelling ability of their students. The scales do not measure the ability to spell the related technical words in industrial education.

[10] Hudelson, Earl, Stetson, F. L., and Woodyard, Ella, *Sixteen Spelling Scales Standardized in Sentences for Secondary Schools*, Bureau of Publications, Teachers College, Columbia University, New York City, 1920.

The validity of the scales is based on the second and third thousand most common words as found in a composite list based on four separate vocabulary studies. The words in the scales are so selected and arranged that each word is one-tenth of a sigma unit more difficult than the preceding word. In administering the tests the words are given in sentences, but the pupils are required to write only the one word that is to be tested. The reliability is reported as being high. Norms are provided for Grades VII to XII.

2. Simmons-Bixler Standard High-School Spelling Scale Forms I, II, III, IV [11]

This unusually valuable spelling scale for high-school use is based upon an extensive program of investigation in high-school spelling undertaken by Mr. Simmons, and supplemented by a revision of the original material under the direction of Dr. Bixler. The result is a series of four forms of scales each containing a preliminary spelling test of 100 words, and 64 scaled lessons of 40 words each. The source of the vocabulary is the socially significant list of words comprising Horn's *Basic Writing Vocabulary; 10,000 Words Most Commonly Used in Writing* [12] after the elimination of a number of abbreviations, irregular forms, and a group of words spelled correctly by 90 per cent or more of high-school freshmen. The scale location of each word is based upon 200 spellings per grade.

In addition to the scaled tests, an alphabetical list of 2910 words is presented with the percentile placement of each word for students in Grades IX to XII. Such a scale constitutes a valuable source of instructional material for use in conducting the spelling "hospital" as well as a useful source of test material of known difficulty.

56. Writing.

The development of the initial skills in writing is one of the functions of the elementary school, but if students are to be legible writers after the formal education period they must be checked continuously by the teachers in all subjects. Although it is desirable that the basic writing habits become more or less automatic, it is also desirable that conscious writing be perfected to such a degree that it will still be legible and of good quality when it does become automatic. Industrial education teachers can aid in the development of good writing

[11] Simmons, Ernest P., and Bixler, H. H., *A Standard High School Spelling Scale*, Turner E. Smith and Company, Atlanta, 1928.

[12] Horn, Ernest, *A Basic Writing Vocabulary*, State University of Iowa Monographs in Education, Series 1, No. 4, University of Iowa, Iowa City, 1926.

habits by demanding legible writing and printing from the students and by displaying high-quality specimens of such work on the bulletin board or on wall charts.

Writing has two characteristics which are important in rating, namely, *quality* and *speed*. Quality is usually determined by having samples of handwriting rated by qualified judges and the samples placed in order of merit on a linear scale. Samples of pupils' writing obtained under standard conditions may then be compared and rated according to the value of the specimen on the scale which it most nearly resembles in quality. Speed is determined by counting the number of letters of standard copy written in one minute. Eighty letters per minute is considered a satisfactory speed for pupils in the ninth grade.

Speed and quality are not rated together. If a pupil of average writing ability writes slowly and carefully the quality of his writing may improve. If he writes very rapidly there is likely to be a reduction in quality. Both speed and quality can be improved through practice. If a pupil is to reach a maximum speed and quality he must also have a good technique (proper position at desk, hold pencil or pen and paper in correct position, etc.).

The rating of handwriting is valuable to industrial education teachers from another angle since it is similar to the rating of quality of workmanship on industrial education projects. It is almost identical with the rating of lettering in drawing, and it has many factors in common with the rating of soldering, riveting, boring, and splicing wire. It is also well to note at this point that speed and quality are measured as separate items. This is also true of speed and quality in rating the results of manual operations in industrial education.

Two handwriting scales which should prove valuable to teachers in rating quality and in diagnosing faults in handwriting have been selected for description. A copy of these scales might well be posted in the shop and used by students to check and analyze samples of their handwriting and as constant reminders to improve their own writing.

1. Ayres Handwriting Scale [13]

The *Ayres Handwriting Scale* now in most common use is known as the "Gettysburg Edition" because the samples in the scales are based upon copy from the first four sentences of Lincoln's "Gettysburg Address." The scale consists of nine widely varying specimens of handwriting graduated by tens from twenty to ninety. Each section

[13] Russell Sage Foundation, New York, New York, 1912.

on the scale is represented by a twelve-line section from the "Gettysburg Address." The relative merit of the specimens was determined by the differences in the lengths of time required by trained judges to read each sample. Thus *legibility* becomes the criterion of merit. This procedure is distinctly in contrast with that used by Thorndike[14] in the development of his *Handwriting Scale*. The results of the use of the two types of scales in the classroom are quite similar, however, in spite of the differences in their construction. Available standards for the various handwriting scales are established only for the elementary-school grades and accordingly are of little value above the eighth grade. However, it may be useful to point out that the writing of junior- or senior-high-school pupils should be quality 60 or above on the *Ayres Scale* at a speed of approximately 80 letters per minute.

2. Freeman's Diagnostic Handwriting Scale[15]

No discussion of measurement of handwriting would be complete without at least a brief mention of the *Freeman Chart for the Diagnosis of Handwriting Faults*. By the use of this analytical chart, attention may be focused upon such qualities as uniformity of slant, uniformity of alignment, quality of line, letter formation, and letter and word spacing. Slant of letters may be revealed by drawing lines through the letter indicating their slant. If the lines are not parallel the lack of uniformity in letter slant is revealed. Alignment may be shown by drawing lines parallel with the bottom and tops of the smaller letters. Weaknesses in letter formation are more difficult to reveal and to classify since there are so many different types. Improperly closed *a*'s and *o*'s and badly formed *n*'s and *u*'s are common types of letter-formation difficulties. Too crowded as well as too widely spaced letters and words operate to reduce the quality of writing. The critical and ambitious teacher of industrial arts subjects will find many opportunities to use this effective analytical scale in bringing about distinct improvements in the handwriting of his students.

57. Measurement of Mathematics.

Throughout the work in industrial arts subjects, frequent demands are made on certain basic mathematical skills. In general, these skills are presented for initial learning in the courses in arithmetic, algebra, and plane geometry.

[14] Thorndike, Edward L., "Handwriting," *Teachers College Record*, Vol. 11: 1-93, March, 1910.

[15] Freeman, F. N., *Freeman Chart for Diagnosing Faults in Handwriting*, Houghton Mifflin Company, Boston, 1914.

Arithmetic. Although arithmetic is rightly considered an elementary-school subject, it is an important factor in achievement in many secondary-school subjects. Arithmetical skills are in demand in practically all classroom and shop activities in the industrial arts. Accuracy in making calculations in connection with shop work and other industrial subjects is an important factor in such achievement. Among the arithmetic tests which are most likely to be of use to the teacher of industrial education are such tests as the *Compass Survey Tests, Advanced Examination,*[16] for Grades IV to VIII, the *New Stanford Achievement Arithmetic Test*[17] for Grades IV to IX, the *Otis Reasoning Tests in Arithmetic*[18] for Grades IV to IX, and possibly certain selected narrow function tests in arithmetic such as the *Compass Diagnostic Tests.*

High School Mathematics. Algebra and plane geometry represent the phases of high-school mathematics of most interest to the teacher of the industrial arts subjects. Among the first-year algebra tests which may readily be of use to the shop teacher is the *Columbia Research Bureau Algebra Test.*[19] In its present form this test is in two parts. Part I is designed to cover the algebra commonly taught in the first semester of the course. Part II covers the second semester's work. A much more intensive type of measurement is provided by the *Iowa Unit-Achievement Tests in Algebra.*[20] These tests are in two equal forms each made up of six tests covering the entire year's work in first-year algebra. The standards represent achievement as of the time when the original instruction on the material was completed.

Achievement in plane geometry may be effectively surveyed by such end-of-the-year tests as the *Schorling-Sanford Plane Geometry Tests*[21] or the *Columbia Research Bureau Plane Geometry Tests.*[22]

[16] Greene, H. A., Knight, F. B., Ruch, G. M., and Studebaker, J. W., *The Compass Survey Tests,* Scott, Foresman and Company, Chicago, 1927.

[17] Ruch, G. M., Terman, L. M., and Kelley, T. L., *The New Stanford Achievement Tests,* World Book Company, Yonkers, New York.

[18] Otis, Arthur S., *Otis Reasoning Tests in Arithmetic,* World Book Company, Yonkers, New York, 1923.

[19] Otis, Arthur S., and Wood, Ben D., *Columbia Research Bureau Algebra Test,* World Book Company, Yonkers, New York, 1927.

[20] Greene, H. A., and Piper, A. H., *The Iowa Unit-Achievement Tests in First-Year Algebra,* Bureau of Educational Research and Service, Extension Division, University of Iowa, Iowa City, 1931.

[21] Schorling, Raleigh, and Sanford, Vera, *The Schorling-Sanford Plane Geometry Tests,* Teachers College Bureau of Publications, Columbia University, New York City, 1925.

[22] Hawkes, Herbert E., and Wood, Ben D., *Columbia Research Bureau Plane Geometry Test,* World Book Company, Yonkers, New York, 1926.

For periodical measurement of achievement over relatively small sections of plane geometry instruction tests such as the *Lane-Greene Unit-Achievement Tests in Plane Geometry* [23] may be used.

In addition to the tests in arithmetic, algebra, and plane geometry which have been described in this chapter there is a real need for tests of related mathematics to parallel the several industrial education courses in electricity, auto mechanics, sheet-metal working, drawing, and printing. Something similar to the type of inventory measurement secured by the *Kilzer-Kirby Inventory Test for the Mathematics of High-School Physics* [24] is greatly needed in these fields. Hunter [25] has recognized this need and has done some pioneer work by developing short tests in shop mathematics, shop arithmetic, and geometry. These tests are not standardized or long enough to be highly reliable, but they are of some value for measuring the mathematics related to industrial arts and for the suggestions they offer for further development along similar lines.

58. Measurement in Sciences Related to Industrial Arts.

Certain contributions of the high-school sciences are apparent in many of the industrial arts courses. Accordingly complete measurement of achievement in these courses must at least include some attention to the fields of high-school physics, chemistry, and general science. Such survey tests as the *Columbia Research Bureau Physics Test* [26] will be found to be very effective measures of end-of-the-year achievement in physics. In a similar way the *Columbia Research Bureau Chemistry Test* will be an effective survey instrument for use by the industrial arts teacher. General science covers so many different phases of the sciences that without doubt it is one of the most useful fields to survey in any attempt to discover the range of information in the sciences held by industrial arts students. For this purpose one of the most useful tests is the *Ruch-Popenoe General Science Test*.[27] Standards are provided for one-semester and year courses in this subject.

[23] Lane, Ruth, and Greene, H. A., *The Lane-Greene Unit-Achievement Tests in Plane Geometry,* Ginn and Company, Boston, Mass.

[24] Kilzer, L. R., and Kirby, T. J., *Inventory Test for the Mathematics of High-School Physics,* Public School Publishing Company, Bloomington, Illinois, 1929.

[25] Hunter, W. L., *Shop Tests, Series No. 2,* The Manual Arts Press, Peoria, Illinois, 1927.

[26] Farwell, H. W., and Wood, Ben D., *Columbia Research Bureau Physics Test,* World Book Company, Yonkers, 1926.

[27] Ruch, G. M., and Popenoe, H. F., *Ruch-Popenoe General Science Test,* World Book Company, Yonkers, New York, 1923.

SUMMARY

Achievement in industrial education cannot be completely and effectively measured without the use of supplementary educational tests selected from other related fields. Since the language skills as represented by reading, language usage, grammar, spelling, handwriting, vocabulary, and composition abilities are so basic and so fundamental to achievement in industrial education subjects, considerable attention is given to the discussion of tests in these fields. The social importance of the correct use of these language skills is so great that no teacher can afford to relax for one moment in his demands for correctness in the oral and written language habits of his students.

Demands are made on certain of the high-school science courses by many of the units of work in industrial education courses. Accuracy and speed in making certain mathematical calculations in connection with shop work are also desirable accomplishments. Accordingly, the teacher of industrial education will wish to sample somewhat liberally the abilities in these other related fields of educational achievement.

SUMMARY EXERCISES FOR DISCUSSION

1. What educational fields appear to be most closely related to achievement in industrial arts?
2. In what specific ways does achievement in industrial arts and other high-school subjects appear to be related to the ability to read rapidly and well?
3. Catalogue the major language skills which should receive the attention of the teacher in industrial arts subjects?
4. In your judgment, what is the most acceptable basis for the selection of a high-school spelling vocabulary?
5. What is the responsibility of the teacher of industrial arts subjects for satisfactory mastery of spelling and handwriting on the part of his students?
6. Secure a copy of the *Ayres Handwriting Scale* and rate at least a dozen samples of handwriting representing a wide range of quality. After two or three days rate the samples again without reference to the scores previously assigned. On what percentage of samples did your two sets of marks agree within five points on the scale?
7. List a few of the more important arithmetical skills which appear to persist into the high school.
8. Why are there no adequate diagnostic tests in algebra or geometry?
9. What special procedures can you suggest for improving problem solving either in arithmetic, algebra, or in the sciences?
10. Compare two selected algebra or geometry tests showing complete lists of specific skills measured by each.

SELECTED REFERENCES

GREENE, H. A., JORGENSEN, A. N., KELLEY, V. H., Examiner's manual for *Iowa Silent Reading Test, Advanced.* Revised Edition. Yonkers, New York: World Book Company, 1931.

GREENE, H. A., and PIPER, A. H., Examiner's manual for the *Iowa Unit-Achievement Tests in First-Year Algebra.* Iowa City: Bureau of Educational Research and Service, University of Iowa, 1931.

HAGGERTY, M. E., and LAURA C., Manual for Administering and Interpreting *Silent Reading Examination, Sigma 3.* Yonkers, New York: World Book Company, 1920.

HORN, ERNEST, *A Basic Writing Vocabulary.* Iowa City, Iowa: Department of Publications, University of Iowa, 1926.

HUDELSON, EARL; STETSON, F. L.; WOODYARD, ELLA, *Sixteen Spelling Scales Standardized in Sentences for Secondary Schools.* New York: Bureau of Publications, Teachers College, Columbia University, 1920.

HUNTER, W. L., *Shop Tests.* Peoria, Illinois: The Manual Arts Press, 1927.

KIRBY, T. J., *The Kirby Grammar Tests.* Iowa City, Iowa: Bureau of Educational Research and Service, University of Iowa, 1920.

LANE, RUTH E., and GREENE, H. A., Examiner's manual for *Lane-Greene Unit-Achievement Tests in Plane Geometry.* Boston: Ginn and Company, 1929.

PRESSEY, LUELLA C., *Technical Vocabularies of the Public School Subjects.* Bloomington, Illinois: Public School Publishing Company, 1923.

SIMMONS, E. P., and BIXLER, H. H., *A Standard High School Spelling Scale.* Atlanta: Turner E. Smith and Company, 1928.

SMITH, H. L., and WRIGHT, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

WILLING, M. H., *The Willing Composition Scale.* Bloomington, Illinois: Public School Publishing Company, 1918.

WILSON, G. M., Examiner's manual for *Wilson Language Error Test.* Yonkers, New York: World Book Company, 1923.

WILSON, G. M., and HOKE, K. J., *How to Measure* (Revised). New York: The Macmillan Company, 1928.

CHAPTER X

TESTING TECHNIQUES IN INDUSTRIAL EDUCATION

I. TYPES OF OBJECTIVE TEST EXERCISES

Certain types of objective exercises which are useful in constructing tests in industrial education are discussed and evaluated in this chapter. Many different forms of test exercises have been used successfully in other subjects to objectify pupils' responses. Doubtless, new or modified types will be developed to meet future testing needs. Industrial education teachers should become critical of such procedures, and should learn to select and use the types of test exercises best adapted to their instructional materials.

59. Objective Techniques Adapted to Testing in Industrial Education.

In the measurement of technical knowledge the usual types of objective exercises can be used. The measurement of manipulative ability requires types of objective exercises designed specifically for the purpose. Two general types of test exercises are used in measuring information and manipulative ability, namely, (1) the recall type and (2) the identification type. In a recall exercise the pupil is called upon to supply the answer from memory. In the recognition or identification types, the pupil must choose the correct response from several possibilities. The latter type involves the recalling of characteristics and relationships but does not call upon memory for the major items of the exercise.

It would be hopeless to attempt to illustrate all the possible forms of objective exercises which have been used in testing, but several examples are given here which should suggest to the industrial education teacher in his construction of tests ways of meeting his measurement needs in the classroom and shop. The different objective types should be studied carefully so that the teacher can readily select those best adapted to measuring different phases of information and manipulative ability.

60. Classification of Objective Test Exercises.

Objective test exercises of types most likely to be of use in the industrial education classroom and shop fall into the following classifications:

I. Recall.
   A. Simple recall.
   B. Completion exercises with one or more key words omitted.
   C. Completion exercises with answers suggested or controlled.

II. Recognition.
   A. Multiple-response tests.
      1. One correct response.
      2. Multiple-answer exercise with varying degrees of merit.
      3. Multiple answers with one or more correct answers.
   B. True-false exercises.
      1. Yes-no questions.
      2. True-false statements.
      3. Diagram and true-false.
      4. Double true-false statements.
   C. Matching exercises.
      1. Word matching.
      2. Picture matching.
      3. Unbalanced column.
   D. Rearrangement test exercises.
      1. Order of operations.
      2. Classification.

III. Performance.
   A. Quality or accuracy.
   B. Identification of tools and materials.
      1. Simple recognition.
      2. Recognition and analysis.
   C. Technique.
   D. Speed or rate of response.

61. Recall Exercises.

The varied forms of the recall-test exercises have been widely used in test construction in all fields. In an analysis of 375 tests from several instructional fields Conneau [1] found that nearly 30 per cent of all

[1] Ruch, G. M., *The Objective or New-Type Examination,* Scott, Foresman and Company, New York, Chapter VIII, p. 189.

the test exercises were of the completion type. The recall exercises are of real value in testing information in industrial education, but of less value in measuring manipulative activity.

The recall type of test exercise has gained favor as a device for the measurement of information in all fields because it is almost entirely objective when properly constructed and administered. Guessing and chance factors operate very slightly. The recall question is a natural form of questioning and is easily and rapidly scored. An important limitation of the recall type of test item is that it tends to be merely factual in character. The recall exercise also requires a great deal of care in preparation, for the reason that unless the missing clue words are carefully chosen several answers will be possible, which will make the scoring difficult at times and bring in the subjective judgment of the teacher. In constructing a completion test it has been found advisable to have each blank call for a single idea, and to avoid a large number of blanks by omitting only a few key words.

Sample Recall Exercises

Simple Recall

1. *Directions:* Answer each of the following questions with a single word. Write the word on the line after the last word of the question.

1. What oil is used in first-quality outside paint? ......
2. In what year was the Centennial Exposition held in Philadelphia? ......
3. Who introduced the Russian system of manual training into America? ......
4. What liquid is commonly used to thin cabinet varnish? ......

2. *Directions:* After each finishing material write the proper thinner.

1. Shellac ......
2. Varnish ......
3. Paint ......
4. Lacquer ......
5. Enamel ......
6. Kalsomine ......

Completion Exercises

3. *Directions*: The following statements are to be completed by adding one, and only one, word in each blank.

1. Oak is a good cabinet ....
2. The length of a meter is ...... feet ...... inches.
3. Cabinet glue should not be heated above ...... degrees Fahrenheit.
4. Wood should not be ...... across the grain.
5. The surface of a cabinet wood is prepared for finishing by ......, ... and ....

Completion Exercises with Answers Suggested

4. *Directions:* Complete the following sentences by inserting one of the words found in the list on the right of the page. The words are to be used only once.

1. A cabinet scraper is used to ...... a surface.	1. Squareness
2. A try-square is used to test for	2. Smooth
3. A ...... is used for cutting lengthwise of the grain.	3. Mill
4. A ...... file is used to shape the edge of a cabinet scraper.	4. Ripsaw

62. Multiple-Response Exercises.

The multiple-response test is one of the most satisfactory objective test exercises to use in the measurement of information and reasoning. On the average it is somewhat more reliable than the true-false type, but is probably not so reliable as the recall test when the tests are equated for length in terms of the number of items in each. It is fairly easy to score, but not easy to construct.

Guessing is a factor which must be taken into account in every objective test form in which the single correct answer must be selected from two or more suggested responses. In theory, at least, guessing is reduced in multiple-response items by increasing the number of suggested responses. There is a practical limit to this, however, for it soon becomes apparent that it is impossible to select large numbers of equally plausible wrong responses for an item. If an exercise were made up with five responses, three of which were so obvious that they would be eliminated at once by a pupil with only a minimum of information, the test exercise would be no more effective than it would be if it were made as an alternate-response exercise to begin with. As a matter of fact, it would be made less effective by the inclusion of the useless material. The tendency at the present time seems to be in the direction of the three-response type. In any event, it is usually desirable to prepare the exercises with the same number of responses throughout the test, if it is to be corrected for chance by the usual formula. The factor of guessing in objective tests is discussed in more detail in Section 71 of this chapter.

In the multiple-choice form of exercise the pupil indicates the correct response by underlining or checking the answer, or by placing the number of the correct response on a blank at the end of the exercise. The writing of the number of the correct response rather than the response itself reduces the amount of writing required of the student and in general is quite satisfactory and objective.

Sample Multiple-Response Exercises—One Correct Answer

5. *Directions:* Each of the following statements can be correctly completed by one and only one of the numbered expressions. You are to write the number which stands for the correct expression on the line at the right of the exercise.

1. Lacquer is thinned with
   (1) turpentine (2) alcohol (3) amyl acetate (4) mineral oil ......
2. A shellac brush is cleaned with
   (1) turpentine (2) water (3) alcohol (4) gasoline
3. No. "00" sandpaper is coarser than
   (1) No. "0" (2) No. "000" (3) No. 2 (4) No. 5
4. Outside paint is thinned with
   (1) water (2) paint remover (3) linseed oil (4) alcohol

Multiple-Answer Exercises with Answers of Varying Degrees of Merit

6. *Directions:* Underline the one word in the parentheses of each statement which best completes the statement.

1. (Walnut, bass, pine, balsa) is a favorite cabinet wood.
2. Mahogany is used for making (barrels, ships, furniture, fence posts).
3. Cypress is used in making (water tanks, beds, floors, boxes).
4. Red wood grows in (Indiana, Iowa, California, Louisiana).

Multiple-Answer Exercises with One or More Correct Answers

7. *Directions:* Underline all the words in each parenthesis which will make true statements.

1. (Cypress, pine, redwood, elm) $\substack{\text{are}\\\text{is}}$ used in making water tanks.
2. (No. 2, No. "000," No. 4, No. "00") $\substack{\text{are}\\\text{is}}$ fine sandpaper.
3. (Oak, pine, basswood, ebony) $\substack{\text{are}\\\text{is}}$ soft wood.
4. (Maple, walnut, fir, yellow pine) $\substack{\text{are}\\\text{is}}$ good cabinet wood.

63. True-False Exercises.

The true-false or "yes-no" form of test exercise is one of the most popular types for measuring information. The true-false exercise is objective, easy to score, has wide adaptability, permits extensive sampling in short working periods, and if ingeniously devised may be used to measure reasoning as well as memory. However, it is not adapted to the measurement of manipulative skills.

High-quality objective exercises of the true-false type are not so easy to construct as it might at first appear. Only materials which are strictly true or false should be put into true-false exercises. Double negatives and trick questions have no place in true-false questions. They should be stated in simple, direct language. The purpose is to

get an objective measure of the pupil's knowledge, and not to confuse or bewilder him intentionally. Any true-false items that are likely to suggest the correct answer to other items should be widely distributed in the test. If reliable results are desired, true-false statements and other forms of test exercises should not be dictated to the class. Paterson [2] reports that dictating test items to a class tends to reduce the reliability of the test. If possible, separate copies of the test should be prepared for each pupil in the class.

The chief limitions of true-false test exercises are that they are open to the influence of guessing and chance factors, and also that they are rather difficult to construct so that the items will be strictly true or false without being too obvious. Attempts to make them less obvious usually makes them ambiguous. Both these limitations can be overcome to a large extent by the thoughtful test worker if he will take unusual care in the construction of the exercises, and correct for guessing when scoring the test.

Two types of alternate-response test exercises (true-false; yes-no) are recognized—the single and double types. The single true-false statement is the more common type and has either a true or a false statement for each fact measured in the test; the double true-false has both a true and a false statement for each concept in the test, *both* of which must be answered correctly in order for the pupil to score on the pair. The paired or double true-false test is a later form designed to control the effect of chance somewhat more definitely by having forced the pupil to respond to both a true and a false item on each fact or concept. The double true-false test undoubtedly does eliminate chance to some extent, but the two test items which relate to the same point must be so distributed in the test that the pupil cannot make a direct comparison of them. A test made up of 100 paired true-false items has been shown to be more reliable than 100 items stated in the usual alternate form, but it requires the same amount of space and time as would be devoted to an ordinary true-false test of 200 items. The reliability and the apparent validity of measurement resulting from the paired exercise test will be somewhat higher.[3] However, the difficulty of making suitable statements of important or basic items in the subject-matter in paired form (in both true and false form) is very great. Experience in formulating true-false exercises soon makes it apparent that certain subject-matter concepts lend

[2] Paterson, D. G., *Constructing New-Type Examinations,* World Book Company.

[3] Greene, H. A., "A New Correction for Chance in Alternate Response Exercises," *Journal of Educational Research,* 17: 102-107, February, 1928.

themselves to statement in one form much more readily than they do in the other. In the development of ordinary true-false examinations it will be found to be a very good practice to go through the basic facts carefully, writing first the false statements which are to go into the test, and then follow with a like number of true statements, which usually are much more easily stated.

True-false exercises have been criticized by some writers because they were of the opinion that the presentation of false forms had a negative effect on learning. Studies by Remmers and Remmers [4] and by Roberts and Ruch [5] have shown that the negative suggestion effects are not so significant as might be supposed. The evidence indicates that true-false statements are a stimulus to learning and are of real value, although this has not been conclusively proved. The burden of proof now seems to rest with those who believe that the true-false question exerts a negative effect.

Samples of True-False Exercises

True-False Statements

8. *Directions:* Examine each statement below and decide whether it is true or false. If true, underline true; if false, underline false. If an item is too hard, skip it and go on to the next one. *Do not guess.* This test will be corrected for guessing.

1. The carburetor maintains the correct proportion of fuel and air at all speeds. True False
2. The generator on an automobile steps up the primary current into a high tension spark. True False
3. The transmission makes possible different speeds forward and reverse. True False
4. The distributor turns the motor over until it has drawn gas in and compressed it. True False

Diagram and True-False

9. *Directions:* Read the following statements about the drawing (Fig. 6) and mark them true or false by referring to the drawing. If a statement is true, underline the word *true;* if false, underline the word *false. Do not guess.* This test will be corrected for guessing.

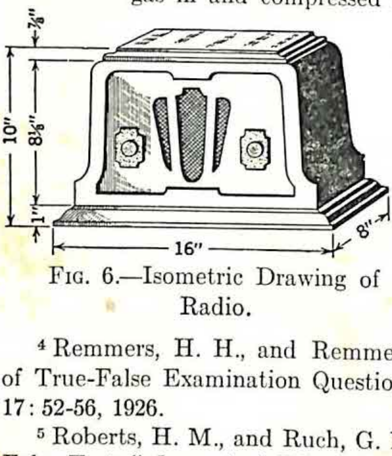


FIG. 6.—Isometric Drawing of Radio.

[4] Remmers, H. H., and Remmers, E. M., "The Negative Suggestion Effect of True-False Examination Questions," *Journal of Educational Psychology,* Vol. 17: 52-56, 1926.

[5] Roberts, H. M., and Ruch, G. M., "The Negative Suggestion Effect of True-False Tests," *Journal of Educational Research,* Vol. 18: 112-116, September, 1928.

1. The radio cabinet is 2′ long.	True	False
2. The top of the cabinet is ⅞″ thick.	True	False
3. The width of the base is 9″.	True	False
4. The depth of the cabinet is 8″.	True	False
5. The base of the cabinet is 1″ thick.	True	False
6. The length of the top is not given.	True	False
7. The front panel of the cabinet is 8″ high.	True	False

64. Matching Exercises.

Matching exercises are valuable in industrial education for measuring relationships between items of information, or tools and materials. The pupil taking the test is called upon to recognize relationships between a test list and an answer list. The pupil usually writes the number of the related item before or after the unnumbered item. Matching tests are objective, easily scored, and in certain subject fields easy to construct. They can be used in measuring factual materials or judgment. In constructing matching exercises, it is important to have 10 or more items to reduce the operation of chance, but if very long lists are used, 25 or 30 items, considerable time is lost by the pupil in sorting the various related pairs. An improvement in this practice is to arrange two separate groups of 10 or more items.

Samples of Matching Tests

Word Matching

10. *Directions:* Below are two columns of words which are related in meaning. Write the numbers of the words in the left-hand column on the blanks in the right-hand column so that they will show the items which are related. For example: "Hammer" is number 1. You find the word "Nail" in the right-hand column. Place the figure 1 in the blank before the word nail. "....1.... Nail." Indicate the relation of all the other items in a similar manner. *Do not guess.*

	....Brace
1. Hammer	....Sheet copper
2. Saw	....Alcohol
3. Mortise	....Linseed oil
4. Open-grain wood	....Sandpaper
5. Turpentine	....Tenon
6. Shellac	....Nail
7. Wood	....Rip
8. Outside paint	....Paste filler
9. Bit	....Varnish
10. Tin snips	....Thumb tack
11. Knife	....Oilstone
12. Drawing board [6]	

[6] Ruch (*The Objective or New-Type Examination*, Scott, Foresman and Company, p. 227) suggests that, when it seems desirable to have less than 10 complete pairs, an excess of statements be made in one column or the other to aid in caring for the chance element.

Parts Identification Test

11. *Directions:* Study the drawing of the automobile engine (Fig. 7). Notice that some of the engine parts are numbered. Below the drawing are the names of the engine parts which are numbered in the drawing. Write the numbers appearing in the drawing in the blanks opposite the correct name for each part.

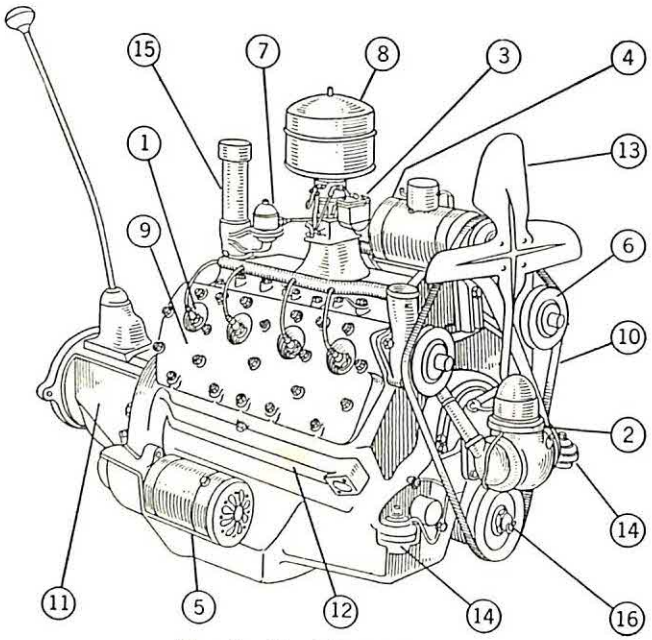


FIG. 7.—Ford V-8 Motor.

Spark plug	....	Gas pump	....
Exhaust manifold	....	Generator	....
Water pump	....	Air cleaner	....
Starter	....	Transmission	....
Carburetor	....	Crank shaft	....
Breather pipe	....	Cylinder head	....
Distributor	....	Fan	....
Fan belt	....	Engine supports	....

65. Rearrangement Exercises.

The rearrangement test is peculiarly adapted to testing information in industrial education which involves chronological order, order of operations, and classification of materials according to grades or quality. It lends itself very well to measuring information which involves the employment of skills in sequence, or as a means of checking in a verbal fashion the plan of a job involving motor skills.

Rearrangement exercises which involve five to eight relations probably do not need correcting for chance. If carefully constructed, rearrangement exercises are objective and easy to score. They are useful in diagnosing special difficulties as they are revealed by a pupil's inability or ability to recognize proper relations in a test situation.

Although this form of test exercise is quite difficult to construct and requires considerable space, it has been used successfully for measuring home mechanics. The following examples from the *Newkirk-Stoddard Home Mechanics Test* [7] serve to illustrate its operation in this field.

Samples of Rearrangement Exercises

12. *Directions:* On this and the following pages are given a number of common jobs in home mechanics. The proper steps for carrying out each job are given here, but these steps are not placed in the correct order. Examine each job in turn, and decide which step should come first. Place the number of this step in the first pair of parentheses, that is, the parentheses at the left. In the same way insert the numbers of the remaining steps in the proper order or sequence, so that when you have finished, one can read the numbers in the parentheses from left to right and find out just how to carry out the steps in the whole job.

Sample: *Job: To Set Casters.*

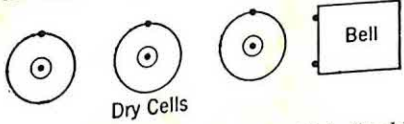
(1) Drive caster-sheafs.
(2) Select a bit and drill the hole.
(3) Select a suitable caster.
(4) Insert the caster and test.
(5) Mark the point for the location of the caster.

Rearrange the numbers to show correct procedure:

(3) (5) ( ) ( ) ( )

13. *Directions:* For the jobs which follow, the connections called for are to be indicated right on the diagrams by drawing in lines with pencil or pen. Read the directions for each job very carefully. When you have figured out how the wires should go, mark them in neatly and clearly. If you don't know all the connections, mark those you think are right.

Job 4. *To Connect Three Dry Cells in Parallel.*
Directions: Show the correct circuit by drawing lines between the black dots.



[7] Newkirk, L. V., and Stoddard, George D., *Newkirk-Stoddard Home Mechanics Test*, Bureau of Educational Research and Service, State University of Iowa, Iowa City, Iowa, 1928.

II. PERFORMANCE TEST EXERCISES

66. Objective Performance Exercises.

Objective exercises for the measurement of performance have not been developed to the point of perfection that characterizes many of the objective pencil-and-paper tests of information. There is need for much additional experimentation to evaluate the most useful types of exercises for the measurement of performance.

Such performance exercises as have been developed are predominantly of the recognition type. The object is to allow pupils to modify materials with tools or instruments or to recognize types or qualities of materials and to check the responses in an objective manner. Objective performance exercises must tell the pupil exactly what to do and not allow the order of major steps to depend on recall. For example, if it is desired to measure a pupil's ability to bore a hole in 1-inch stock with a No. 6 bit, he should be given the bit, brace, and stock with directions, but should be carefully supervised to see that the directions are followed. This will result in a sample of the pupil's work under standard conditions which can be rated and compared with similar samples of other pupils' work. If, on the other hand, a pupil is told to get a No. 8 bit and to bore the hole and he uses a No. 16 bit with which to bore the hole, the sample will not be entirely comparable with the No. 8 samples. If the pupil makes a mistake in the selection of the specified size of bit, it may indicate in a rough way that he does not know much about the sizes of bits, but it adds an uncontrolled variable to the performance factor without completely testing the pupil's knowledge of the sizes of bits. Knowledge of the different sizes of bits and ability to bore a hole are two different things, from the standpoint of test construction. The situation is similar to that in which a teacher asks a pupil to write down the name of a cabinet wood, and the pupil writes the word "wallnut." The pupil's response is correct, but his spelling is faulty. The temptation is to lower the grade because of a misspelled word although the answer is correct. In this case the pupil should have a mark in spelling and a mark for his knowledge of the wood, but these two variables should not be allowed to interfere with each other. The same is true of the sizes of bits and the ability to bore a hole. They are independent variables both of which should be tested, but not in the same type of situation.

Performance test exercises may be divided into four groups according to use, namely, (1) tests of quality or accuracy, (2) identifica-

tion of materials or tools, (3) technique of tools or instruments, and (4) speed or rate of response. These four concepts, which have been discussed in Chapter V, will be briefly reviewed here with illustrations of test exercises which have been used in measuring the factors.

67. Quality or Accuracy Exercises.

The quality or accuracy of industrial education work is determined by carefully evaluating materials which have been modified in some significant way with tools, materials, or instruments. A test exercise for measuring quality of workmanship must allow the pupil to modify materials under genuine and controlled shop conditions, so that the results can be rated with reasonable objectivity and compared with the results of other pupils who have practically the same background and physiological development. The pupil not only must modify materials under controlled conditions, but also must modify enough materials to give an adequate or reliable sampling of the abilities being measured.

The following guiding principles may be helpful in constructing objective test exercises of quality:

1. Provide a job which will give adequate samples of the results of the tool or instrument operations being measured.
2. Give specific directions for doing the work.
3. Provide all tools and materials necessary.
4. Measure the results by physical measurements, quality rating scales, and where necessary, by inspection.

SAMPLE OF QUALITY OR ACCURACY EXERCISES

14. *Operation:* To saw to a line with an 8-point cross-cut saw. Saw as accurately as you can.

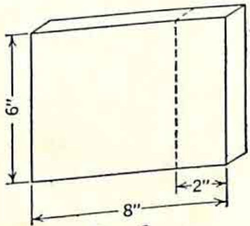


FIG. 8.

Materials: Eight-point cross-cut saw in good condition. A soft wood board free from knots, 7/8" x 6" x 2', surfaced on four sides and laid off as shown in the drawing.

Directions: Place board in position for sawing.

15. *Operation:* To bore a hole with a ½″ wood bit.

Tools and Materials: A ½″ bit and brace, a piece of soft wood as indicated in the diagram, a bench, vise, and try-square.

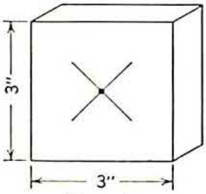


FIG. 9.

Directions: Bore holes through the wood block perpendicular to the surface at the point indicated.

16. *Operation:* To cut a line using tin snips.

Tools and Materials: A pair of sharp, properly adjusted tin snips, a piece of XX tin plate as shown in the diagram with the cutting line marked.

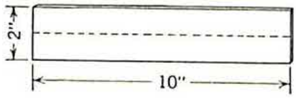


FIG. 10.

Directions: Cut the tin into two 1″ strips. Cut directly on the line.

Examples 14, 15, and 16 are samples of short manipulative test exercises designed to measure ability to saw to a line with a cross-cut saw, ability to bore a hole perpendicular to a surface with an auger bit, and ability to cut strips of tin with tin snips. Many tool operations can be tested in this manner. The construction of complete tests of quality or accuracy is discussed in detail in Chapter XI.

68. Identification Exercises.

Identification exercises are very useful for testing the pupil's ability to recognize materials, instruments, and tools. They are also used for measuring a pupil's ability to analyze special difficulties. The following significant principles in the construction of identification exercises should prove suggestive to the teacher:

1. Provide a representative sample of the objectives to be identified.
2. Suspend materials so that they can readily be examined.
3. Score the items by checking the objective written responses.

The identification exercise is easy to use and is objective in scoring, and the same sample panel can be used by changing for testing the pupil's ability to identify a number of different materials, fixtures, or tools. The authors have found it advantageous to suspend the items

because it allows the pupil to hold the items in his hands, to lift them, to smell them, etc. This gives a natural psychological approach. It also prevents certain optical illusions. For example, it is difficult to realize that a 6 penny nail is not a 3 penny common when it is fastened securely beside a 60 penny spike.

Samples of Identification Exercises

Identification of Materials

17. *Directions:* Number your paper from 1 through 8 along the left-hand margin. Opposite each number write the name of the wood that is hung under the corresponding number on the panel.

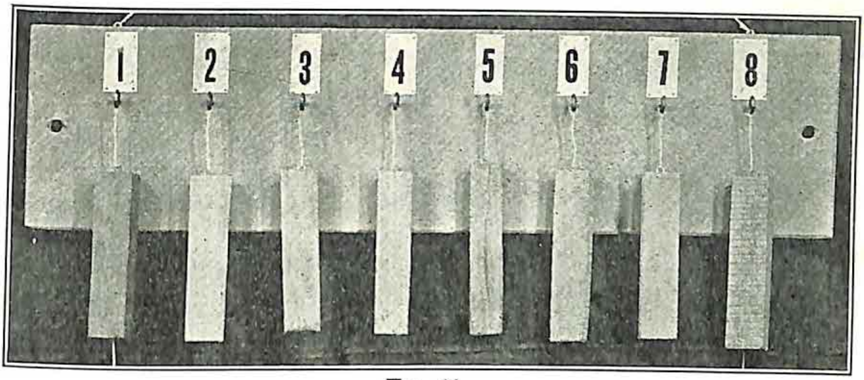


Fig. 11.

Analysis and Identification of Defects in Bells and Buzzers

18. *Directions:* Number your paper from 1 through 6 along the left-hand margin. Opposite each number write any defect in the bell or buzzer that is hung under the corresponding number on the panel. *Do not guess.*

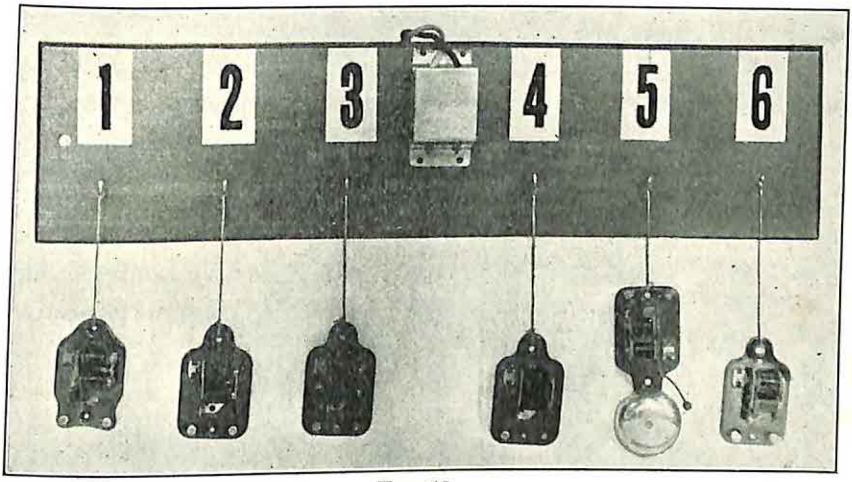


Fig. 12.

Testing for Defective Fuses

19. *Directions:* Number your paper from 1 through 6 along the left-hand margin. Test out the fuses with a test lamp. Opposite each number on the paper indicate whether the corresponding fuse on the panel is blown or satisfactory. *Do not guess.*

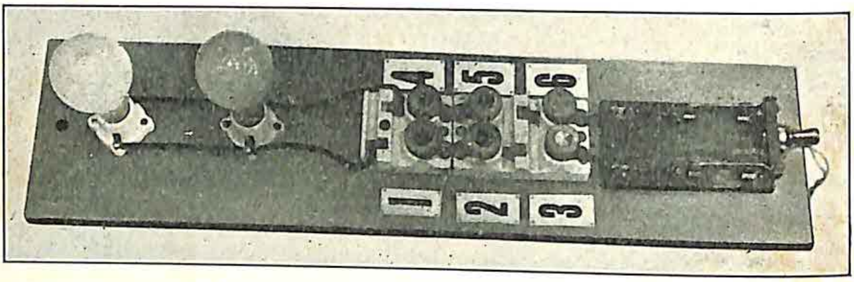


FIG. 13.

69. Technique Exercises.

Technique exercises are designed to measure a pupil's method of manipulating tools, machines, instruments, or materials. It is possible to do work of good quality with poor technique, but great skill can scarcely be developed without the fundamental techniques with tools and materials. Shop technique does not lend itself to measurement with complete objectivity. The technique exercises require the pupil to do certain things which demand the manipulation of tools and materials, which then provide a means of rating the major techniques. Test exercises of technique, like exercises of quality, require thought and experimentation in their construction but are valuable in measurement, diagnosis, and teaching.

The teacher will do well to consider the following guiding principles in the construction of test exercises for measuring technique:

1. Provide activities which will call for the use of tools or instruments in which technique is to be rated.
2. Give specific directions for doing the work.
3. Provide enough activity to give adequate samples of the various techniques.
4. Provide necessary tools and materials.
5. Rate the techniques by using a rating scale.

TEST EXERCISE ON TECHNIQUE

20. *Operation:* To saw to a line.
Directions: Saw the board on the line as marked, perpendicular to the surface.

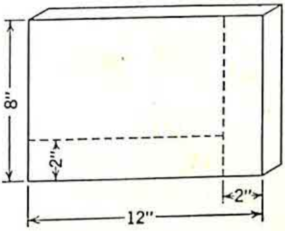


FIG. 14.

Tools and Materials: Ripsaw and cross-cut saw in good condition. Bench, vise, and a piece of soft wood marked as indicated in the drawing.

Rating Scale.[8] As the pupil makes the cuts with the saw, the following points are observed and checked. Each item is rated on the basis of 10, and the score is determined by adding the ratings.

Sawing

1. Clamping stock. I I I I I I I I I

   Stock should be held so that it will not be loosened or cracked and should also facilitate sawing.

2. Starting cut. I I I I I I I I I

   With thumb at line, saw should be placed against the thumb. Saw should be pulled back slowly a few times to make a groove, then pushed forward.

3. Holding saw. I I I I I I I I I

   Saw should be held in right hand. For cross-cut, angle should be 45 degrees; for rip, 60 degrees.

4. Stroke I I I I I I I I I

   Stroke should be long and even, not too fast. Proper angle should be kept during sawing. Line should be followed.

5. Ending cut. I I I I I I I I I

   One should reach over with the left hand and hold on to the piece being cut off. Saw strokes should be slow with little pressure to prevent breaking off the end.

70. Speed or Rate of Response Exercises.

Rate of response is of considerable value in trade courses, but of less importance in the cultural courses of the elementary and junior

[8] Sample 20 gives the method of rating a shop technique. The construction of tests for rating techniques is discussed in more detail in Chapter XI.

high school. Tests of speed will be discussed and illustrated in detail in Chapter XI, but it seems desirable to point out here that quality must be clearly defined and held nearly constant or rate of response cannot be measured accurately. Speed and accuracy are each variable factors in achievement and performance. Test exercises designed to measure speed or rate of response must present a well-defined activity with appropriate standards. When the pupil can do the activity and meet the required standards, then he is ready to take the test to see how rapidly he can do the problem. Practice will result in a pupil's improving his score up to the point where further increase is limited by native ability. The amount of speed a pupil should have can be determined by the demands of the job or by comparison with the best efforts of others.

In the construction of exercises designed to measure rate of response the following principles should be observed:

1. Define the exact work to be done.
2. Give definite standards.
3. Give the pupil a chance to achieve the standards and learn exactly what they are.
4. Do the job on a carefully controlled time basis.
5. Score on the basis of known quality and time.

Sample Rate of Response Exercise

21. *Operation:* Rate of boring holes through ⅞″ soft wood.

Part I

Preliminary Activity: Bore three ½″ holes through the piece of soft wood given you at the points indicated in the drawing and on the board.

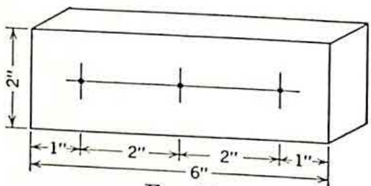


Fig. 15.

Practice boring holes until you can do as well as or better than the sample given you by the teacher.

Part II

1. Say to the pupil, "Now that you can bore holes accurately and neatly, we want to learn how quickly you can bore three holes. Do your work the same way you did in practicing. Continue to bore holes which are as good as or better than the sample."

2. Be sure that the pupil has a piece of wood ready for boring and that the bit is in the brace and that the vise is in working order.

3. Say to the pupil, "Ready, begin." Check the amount of time in seconds that is required to bore the holes.

4. Score the exercises by retaining each sample of boring that is as good as or better than the minimum sample. For example, the pupil bores the three holes in 60 seconds. Two holes are satisfactory, but one is inferior, so the pupil's score on the exercise is two holes in sixty seconds.

III. CHANCE FACTORS IN OBJECTIVE EXAMINATIONS

71. Guessing in Objective Tests.

Test exercises of the recognition type, in which one or more suggested wrong responses accompany the correct response, are definitely affected by the factor of chance or guessing. Ordinary recall items which call upon the student to initiate and state his response naturally are not influenced by this factor. Most alternate-response (true-false; yes-no) items open up the possibility of a fifty-fifty chance of the individual's guessing the correct answer in all items about which he has no information at all. Multiple-response items of the three-, four-, or five-response types decrease this probability as the number of alternate responses is increased. Within certain limits, chance operates in matching exercises, and to a smaller degree in exercises using the rearrangement or the classification testing techniques.

The actual degree to which chance affects a pupil's score is almost impossible to determine. It depends upon the form of the test exercises and their arrangement in the test. It also depends upon the amount of information, or lack of it, which the pupil has concerning the specific item. If we reason from an *a priori* basis, it is quite apparent that the pupil who is totally ignorant of the facts involved in a test item has a fifty-fifty chance of guessing the correct response in a true-false or two-response test. For instance, if an individual were to respond to a properly balanced true-false test with the exercises themselves covered with a sheet of paper, by marking at random the true-false responses along the margin of the paper, this would represent a situation in which total ignorance of the items actually operated. If the test were long enough to provide a reasonable sampling, the resulting score on the test under these conditions should be zero, since no knowledge of the test content would be called into play in responding to it. Pure chance would be operating. Under these conditions the individual should mark almost exactly the same number of wrong responses as right ones. If the number of right and wrong

responses did not check closely, it would mean that the test itself was not properly balanced, or that it was not long enough to be reliable.

Now in actual practice, there should be very few items in an examination about which the student should be totally ignorant and be forced to resort to pure guessing, if the test is made up of valid items, selected from the actual content which the student has had an opportunity to learn. Accordingly, this fringe of knowledge, slight though it may be, should enable him to succeed more often than he fails. In other words, pure guessing does not operate in the use of a valid alternate-response test. In theory, guessing would operate to increase the score through a lucky guess just as often as it would tend to reduce the score through an unlucky guess. The fact that the pupil is never so ignorant of such a test item as this reasoning would assume, makes it desirable to conclude that he at least guesses one exercise right for each one guessed wrong. This results in reducing his score in the number of exercises right by the number of exercises he missed in the test.

The best available evidence seems to indicate that the apparent validity of alternate-response tests is increased slightly by assuming that guessing actually does take place and correcting the score on that basis. The net result of the application of this type of correction is possibly to over-correct slightly, but in most cases this is not serious. Ruch [9] has suggested that if correcting for chance in recognition forms appears to be unsatisfactory, approximately the same effect may be brought about through increasing the length of the test by the use of 10 to 15 per cent more test items than would be required for the expected reliability of measurement.

72. Correcting for Chance in Objective Tests.

The typical procedure for the correction of exercises for the operation of chance may be generalized in the following formula:

$$C = R - \frac{W}{N - 1}$$

in which C is the corrected score, R is the number of exercises answered correctly, W is the number of exercises answered incorrectly, and N is the number of choices in the exercise. Thus, if there are 5 choices in a multiple-response examination, the correction consists in taking $\frac{1}{4}$ of the number of wrong answers from the number of exercises answered correctly. In true-false or other alternate-response tests, the formula works equally well. In true-false tests N equals 2.

[9] Ruch, G. M., *The Objective or New-type Examination*, Scott, Foresman and Company, Chicago, 1929.

Thus, the denominator of the formula becomes 1, and the net result is to deduct the number of wrong responses from the number answered correctly. For example, a student in responding to a true-false examination consisting of 125 items, omits 11 and answers 14 incorrectly. The number of exercises he answered correctly (R) is found by subtracting the omitted and incorrectly answered exercises from the total number of items in the test. $125 - 11 = 114$; $114 - 14 = 100$, the number right. The correction for guessing involves taking the wrongs from the rights ($R-W$). Accordingly, the corrected score is $100-14$, or 86. In cases where the student misses more exercises than he answers correctly, the practice ordinarily followed is to assign scores of zero, rather than to show a negative score. Practically, the individual could scarcely know less than zero, and furthermore, it is likely that such a situation arises out of the unreliability of the test itself.

Attention should possibly be directed once more to the matter of the specific instructions to be given the student in the use of recognition-type tests. The best practice, based on a conservative estimate of the available evidence, seems to be to *direct the pupils not to guess* in taking the test, *but to correct the resulting scores* on the test exactly *as if they had guessed*. The only exception to this general rule seems to arise in the use of double true-false exercises. In this case, it appears desirable to encourage the student to attempt to answer every possible exercise in both parts of the test. The method of scoring the test in terms of pairs right takes adequate care of any tendency or necessity on his part to resort to pure guessing. Furthermore, unless the items are utterly invalid, he must have a fringe of information about many items which he might be tempted to omit under the conditions of the typical true-false test. Since missing or omitting *one or both* of the paired exercises makes it impossible for the pupil to score on that pair, he should be given the benefit of the doubt and a chance to score on every pair of exercises.

IV. TYPES OF ESSAY-TYPE EXAMINATION EXERCISES

Although a general program of measurement of classroom products in the industrial arts is most likely to be advanced through the elimination of the subjective features of the teacher's judgment at all possible points, it must nevertheless be recognized that certain desirable products of the classroom and shop simply do not lend themselves to the objective approach. Furthermore, many teachers of industrial education wish to make use of essay-type tests occasionally for other reasons. Since this type of test is still used, and probably

always will be to some extent, it is unquestionably desirable to point out here some of the possibilities and limitations of this type of measurement, as well as to submit certain suggestions which if carefully utilized may result in the distinct improvement of the less objective methods of measurement.

Essay-type examinations, though generally not so reliable as the average objective examination, frequently secure measures which are just as valid as they would be if stated in objective form. The lack of reliability in the essay or traditional examination lies mainly in the limited extent of the sampling which the use of this form of question permits, and in the lack of objectivity in scoring the items. Many examinations composed entirely of essay questions are valid in the general sense of the term. Their limitation results from the incompleteness of the sampling taken and from the uncertainty with which the results are evaluated.

73. Traditional Examination Questions.

The traditional or discussion-type examination is almost uniformly made up of recall questions. The following types are representative:

I. SIMPLE RECALL.

Samples:

1. Name four different types of wood stains.
2. Name the different grades of sandpaper used in woodfinishing.
3. Name the ingredients of paste wood filler.

II. DESCRIPTION.

Samples:

1. Why is walnut a good cabinet wood?
2. What are the chief characteristics of red wood?
3. Why is balsa a favored wood for constructing model air craft?
4. What are the characteristics of quarter-sawed oak?

III. COMPARISON AND ANALYSIS.

Samples:

1. What is the difference between a superheterodyne circuit and a radio-frequency circuit?
2. How does a dynamic speaker differ from a magnetic one?
3. What is the difference between an inside aerial and an outside aerial?
4. How does a battery set differ from an a-c set?
5. Is there a difference in the underlying principle of head phones and a magnetic speaker? Explain.

IV. Procedure.

Samples:

1. Give the steps in applying a rubbed varnish finish.
2. Give the steps in squaring a board.
3. Give the procedure for fuming oak.
4. Give the procedure for tinning and soldering copper.

74. Constructing Essay-Type Exercises.

On first thought the essay-test exercise seems easier to prepare and use than the objective type. In the way the traditional examination is ordinarily used, or perhaps we should say misused, it does take less time to prepare and the results obtained are much more subjective and unreliable than those from objective tests. If the essay-type test is constructed so as to give a fairly reliable result it is not easier to construct than the objective test exercise. In fact, it may demand a great deal more time and careful thought.

The following rules have been found very helpful in the construction of essay-type questions.

1. State the question in a simple, direct manner so that it demands the reproduction, comparison, or evaluation of a specific unit of instructional material.

Example: *Poor.* Name all the stains you can.
Better. Name four types of wood stains.

2. Write out just exactly the answer that is expected for each essay question in the test. This may be either in outline, brief paragraph, or diagram.

Examples: 1. Name four types of wood stains.

Teacher's answer: 1. Water stains.
2. Oil stains.
3. Spirit stains.
4. Chemical stains.

2. Why is walnut better suited to cabinet work than fir?

Teacher's answer: Walnut is better suited for cabinet work because of its natural beauty, color of the wood, close grain, and durability; it is stronger than fir, does not splinter as readily and takes a better finish.

75. Scoring Essay-Type Test Exercises.

The more objectively the essay test exercises can be scored, the less the results will be influenced by the personal judgment of the scorer. The following suggestions have been found valuable for use in correcting essay-type exercises:

1. Tests should be scored by the one who makes out the questions. He should know exactly what responses are intended and write them down.

2. Each pupil taking the test should write his name on the back of the test paper, and the scorer should disregard the name until the test is scored. This eliminates the subjective factor of being influenced or biased in judgment because of former contacts with the pupil.

3. The scorer should not mark off for misspelled words, sentence structure, paragraphing, poor writing, etc. Similarly, he should not increase the score for excellence in these things. However, such factors may be indicated or checked on the examination. The reason for this is that the test is to measure the pupil's knowledge of certain information in an industrial education course. If it is desirable to test a pupil's ability to write, spell, or use correct written English, suitable tests should be given for this purpose which are valid and reliable.

4. Essay test exercises can be corrected most simply by correcting each item in all the tests rather than by correcting the entire tests separately. This enables the scorer to concentrate on the answer to one test exercise and thus he is better able to judge the merits of the several pupil responses to the same question.

5. Rate each question on a scale of 10 or 20 and then add the ratings on all the essay test exercises to get the mark for the paper. This method helps to objectify the score. The score is based on a number of careful judgments rather than on one complex judgment for the entire score.

The essay-type exercise can be made much more objective and the subjectivity of the teacher's marks can be significantly reduced if a method similar to the one outlined in the preceding paragraphs is followed. However, certain limitations of the essay-type question are obvious. Frequently the time gained by the teacher in preparation is lost in scoring. At best, the essay-type test is not as valid or reliable as an average objective-type test. Research studies have shown the reliability to be around .59 on an average.

Kelly [10] and Fauber and Ruch [11] have shown that subjectivity of teachers' marks can be reduced significantly through the use of scoring rules. But even if the subjectivity of teachers' marks could be reduced by half, they still would not provide measures which are nearly so reliable as those obtained from objective tests. Ruch states that "Experience and experiment have shown that the results of an essay examination cannot be evaluated fairly by human minds." [12] In addi-

[10] Kelly, F. J., *Teachers' Marks,* Teachers College Contribution to Education, No. 66, p. 83, Columbia University, New York, 1914.

[11] Unpublished master's thesis, 1926, University of Iowa.

[12] Ruch, G. M., *The Objective or New-Type Examination,* Chapter I, p. 20, Scott, Foresman and Company, Chicago, 1929.

tion to the psychological difficulty of making a complex judgment, there is also the serious disadvantage of limited sampling, which has been mentioned previously. The average objective tests will sample at least five times as widely into a field of information as an essay examination requiring the same testing time.

SUMMARY

The more important techniques for use in the construction of traditional examinations and informal objective tests are summarized in this chapter. The differences between the standardized test and the informal objective examinations are pointed out.

The problems involved in controlling the chance or guessing factor in certain forms of objective tests are treated briefly in this chapter because of the close relation of this factor to the technique of reliability of measurement used. Recognition is given to the fact that not all the measurement that goes on in the classroom and shop should be objective.

SUMMARY EXERCISES FOR DISCUSSION

1. Why are many of the paper-and-pencil tests which are useful in other educational fields not well suited to the demands of objective measurement in industrial subjects?
2. Illustrate by example each of the main types of objective exercises suited for use in measurement in industrial arts courses.
3. What are the main advantages and disadvantages of objective examinations?
4. Show how the general formula for correcting for guessing in objective tests actually works in an alternate-response test and in a five-response test.
5. Evaluate the suggestions for improving the objectivity of scoring of essay-type exercises.

SELECTED REFERENCES

BUCKINGHAM, B. R., *Research for Teachers.* New York: Silver, Burdett and Company, 1926.

FOSTER, R. R., and RUCH, G. M., "On Corrections for Chance in Multiple-Response Tests," *Journal of Educational Psychology,* Vol. 18: 48-51, 1927.

GREENE, CHARLES E., "New Type Tests," Research Monograph No. 3, Denver, Colorado, 1926.

GREENE, H. A., and JORGENSEN, A. N., *The Use and Interpretation of Educational School Tests.* New York: Longmans, Green and Company, 1929.

LANG, ALBERT R., *Modern Methods in Written Examinations.* Boston: Houghton Mifflin Company, 1930.

ODELL, C. W., *Traditional Examinations and New-type Tests.* New York: The Century Company, 1928.

ODELL, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

ORLEANS, J. S., and SEELY, G. A., *Objective Tests.* Yonkers: World Book Company, 1928.

RUCH, G. M., *The Objective or New-type Examination.* Chicago: Scott, Foresman and Company, 1929.

SMITH, H. L., and WRIGHT, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

SYMONDS, P. M., *Measurement in Secondary Education.* New York: The Macmillan Company, 1927.

TOOPS, H. A., "Trade Tests in Education," Teachers College, Columbia University, Contributions to Education, No. 115, 1921.

WEIDEMANN, C. C., "How to Construct the True-False Examination," Teachers College, Columbia University, Contributions to Education, No. 225, 1926.

WILSON, G. M., and HOKE, K. J., *How to Measure.* New York: The Macmillan Company, 1928.

WOOD, BEN D., *Measurement in Higher Education.* Yonkers, New York: World Book Company, 1923.

CHAPTER XI

CONSTRUCTION AND USE OF INFORMAL SHOP TESTS

I. CONSTRUCTION OF AN INFORMAL OBJECTIVE TEST

76. Steps in Building an Objective Information Test.

In contrast with certain other school subjects, the industrial arts subjects present two widely different phases of achievement for measurement. One of these phases is expressed in terms of the ability of the individual student to go into the shop, and, by following specific directions, come out with a product of a given quality which is itself evidence of achievement. This is a test of performance. The other is expressed in terms of knowledge of facts and their relationships which may lie back of the student's actual performance. This is a test of information. The performance test calls for direction, action, production. The information type is usually a paper-and-pencil test. It is obvious that the best possible test of information cannot be wholly valid for any industrial education course because of the fact that it deals only with information and does not measure such factors as quality or rate of response, techniques, and personality traits, which unquestionably are important elements of the course. Both are essential to complete measurement of accomplishment in this field. The testing techniques for both types of tests are discussed in the preceding chapter. The steps in constructing informational types of tests, and in deriving rating scales for the evaluation of the quality of products obtained under performance testing conditions, are set forth in this chapter.

The distinctive feature of the teacher-made objective examination which makes it especially useful in the evaluation of classroom achievement is the closeness with which its content can be made to parallel the subject-matter actually taught to the class. This is merely another way of stating that its validity is high in proportion to the extent that the teacher includes in the examination exercises sampling from facts which the students have had an opportunity to learn. A teacher who knows his pupils and his subject-matter may readily construct an objective examination which will have all the merits of the standardized test (except the standards or norms them-

selves), with few or none of the limitations. Critical selection of items from the important phases of the subject which have been given instructional emphasis will guarantee the validity of the test. Observance of the simple principles of formulating objective exercises will produce an objective test. A wide sampling over the significant phases of the subject will produce a test long enough in terms of items and working time to secure reliable results. Teacher-made tests which meet these three criteria leave little to be desired.

77. Securing Validity.

Objectivity and reliability in an examination are qualities which are functions of the form of the exercises used and the breadth of sampling of items taken. That is to say, they are the results of the application of certain principles of measurement which can be learned by any classroom teacher. The first constructive step in the development of the informal objective examination is the establishment of a basis for the validation of its content.

Course of Study as the Basis for Validity. The validation of an examination presumes a knowledge of exact details of the curricular material to be taught, as well as a background of experience and judgment adequate for a critical evaluation of the social and practical significance of the various units of instruction. This means, clearly enough, that the teacher must have an intimate knowledge of the content of the course of study.

In order to build up a suitable background for understanding the development of the objective tests presented later in this chapter for purposes of illustration, the following course-of-study outline in woodworking is presented. The content of this outline is not put forward as ideal, but rather as a body of information and skills affording materials suitable for illustrating several types of tests useful to the industrial education teacher. The illustration from woodworking is used here mainly because it is the most widely taught and best understood instructional division in industrial education, and because practically all the testing techniques suitable for use in this field may be applied to other industrial education subjects.

Course-of-Study Outline in Woodworking. This outline is presented as a definite basis for the construction of an objective examination for an eighth-grade class in woodworking. The unit of instruction covers ten weeks of woodworking representing one of four instructional divisions of a course in general shop. The class itself is made up of twenty-six boys coming from middle-class homes typical of a mid-western city.

Objectives of the Course

1. To develop an appreciation of good materials and workmanship.
2. To develop handyman abilities with common tools and materials.
3. To develop hobbies for leisure-time activities.
4. To further intelligent choice of life occupations.
5. To give information about the industries and their workers.
6. To develop desirable social traits and attitudes.
7. To provide opportunity for planning and problem solving.
8. To motivate and vitalize academic learning.

What the boys should be able to do with woodworking tools[1]

1. To use a rule in measuring.
2. To use dividers or compass for laying out curves and dividing spaces.
3. To use a try-square for testing.
4. To adjust a plane.
5. To square a piece of stock.
6. To saw to a line with a rip or cross-cut saw.
7. To use back saw.
8. To use coping saw.
9. To bore holes in wood.
10. To fasten with screws.
11. To trim or pare with a chisel.
12. To use scraper.
13. To use sandpaper.
14. To drive and draw nails.
15. To lay out and cut a chamfer.
16. To glue up work.
17. To fit hinges.
18. To make butt joint.
19. To make dowel joint.
20. To sharpen edge tools.

What the boys should know about wood and the divisions of the industry

1. Know the principal characteristics, working qualities, principal uses, and sources of supply of the following woods: pines, cypress, oak, walnut, ash, birch, maple, mahogany, red cedar, hickory, gum, chestnut, and poplar.
2. How lumber is cut and milled.
3. Standard dimensions of lumber.
4. Knowledge of veneer and plywood.
5. Kinds of glue and its preparation.
6. Kinds of nails and their uses.
7. Kinds and sizes of screws.
8. Kinds and grades of sandpaper.
9. Grades and uses of steel wool.
10. Distinguishing characteristics of period furniture.

[1] Adapted from the A. V. A. Committee's Report on *Standards in Industrial Arts*

11. Basic principles of good design in furniture.
12. Use of common types of hinges and fasteners on woodworking projects.
13. Kinds of grinding and sharpening stones.
14. Location of manufacturing concerns and labor conditions.

What the attitude of the boys should be [2]

1. Industrious.
2. Cooperative.
3. Self-reliant.
4. Considerate of the rights of others.
5. Ready to assume responsibility.
6. Loyal.
7. Fair minded.
8. Optimistic toward life.
9. Law abiding.
10. Appreciative of duty in common things.

Selection of Major Groups of Informational Items. The next step in the validation of the content of an informal objective test of information is to select, in the light of the objectives set up for the course, the groups of skills which are informational in character and which can be measured by means of a paper-and-pencil test. The following summaries represent the major groups of such informational aspects found in the foregoing outline on woodworking:

1. Different types of planes.
2. Different types of saws.
3. Sizes of wood bits.
4. Sizes of screws.
5. Kinds and sizes of chisels.
6. Procedure in squaring stock.
7. Sizes of sandpaper.
8. Sizes of nails.
9. Glue and its use.
10. Different types of hinges.
11. Kinds of wood stain.
12. Types of fillers.
13. Different types of brushes.
14. Composition of shellac.
15. Enamel and its composition.
16. Varnish and its composition.
17. Different kinds of paint.
18. Composition of wax.
19. Composition of lacquer.
20. Common joints.

[2] This unit is the same for all shop subjects, and probably for the entire curriculum.

21. Steps in applying stain, filler, shellac, varnish, enamel, paint, wax, and lacquer.
22. Steps in squaring stock, preparing glue, using sandpaper, sharpening wedge edge tools, boring holes, and fastening with screws.
23. Principal characteristics and uses of common woods.
24. Dimensions of lumber.
25. How lumber is cut and milled.
26. Veneer and plywood.
27. Grades and uses of steel wool.
28. Principles of design.
29. Characteristics of widely known period furniture.
30. Types of grinding and sharpening stones.
31. Manufacturing concerns and labor conditions.

Suggestions for Securing a Valid Sampling of Informational Content. The specific problem of this discussion is to demonstrate how it is possible to secure a valid sampling of the informational content of the course. The following rules will be found very helpful in accomplishing this:

1. Keep clearly in mind the objectives of the course. Try to formulate questions which will measure the extent to which the objectives have been achieved. Emphasize the relative and social utility of the subject-matter and avoid purely factual questions unless they are essential to building up concepts.

2. Ask questions which the objectives indicate are of most importance, but under no circumstances ask questions included merely to "stump" the pupils. Trick questions, and unusually difficult ones, are only dead weight in the test, waste valuable testing time, and in general lower the validity of the test.

3. Ask a large number of questions over all parts of the course. The different types of objective test exercises are best suited for testing a large number of items in the time ordinarily allotted to measurement.

4. Have other teachers make suggestions as to the importance of the exercises selected for the test. Take into consideration the comments of pupils as to the value of the different test items. If the pupils consider them unfair, obscure, and too easy, they should be eliminated or modified before the test is used again.

5. The test cannot be more valid than the course of study on which it is based. The progressive teacher will revise his course and tests from time to time to bring the work abreast with good practice and the results of curriculum research.

In establishing validity it is a good policy to construct 200 or 250 test exercises based on the course. This furnishes sufficient material

to eliminate undesirable test exercises and still have adequate material for at least two forms of the test. One form is sufficient for a valid examination, but two forms make it more valuable. The second form may be used for testing those absent from the first test, or the test may be used from year to year, alternating the two forms.

78. Securing Objectivity.

After the major informational topics have been agreed upon in the light of the teaching objectives and, when possible, passed upon by other teachers, the items can be expanded and developed into objective test exercises. It is a good policy to use the type of objective exercise which best fits the material, rather than to attempt to make all exercises conform to a single type, as true-false, multiple-choice, etc. Chapter X gives examples of test exercises which have been found valuable in testing information in shop courses.

Two fairly satisfactory methods of procedure are suggested for recording the test exercises as they are developed. In one, the teacher may write the exercises on sheets of paper allowing a half inch between each question, so that, after the questions have been formed, the paper can be cut into strips with one question on a strip. These strips can be shifted to eliminate the less desirable ones according to the teacher's plans. Similar types of test exercises can be grouped to save time in manipulation of the items. In the other method, the teacher may use 3 inch by 5 inch cards and a card index. Each question is put on a separate card, and the cards are grouped according to the type of test exercise used. The first draft of the questions in either procedure should be double spaced to allow for corrections by the teacher after he has given them critical analysis. The cards can be shifted or eliminated as desired. The authors have found this second method to be handier and neater but a little more expensive.

After the test items have been developed and classified according to types of questions, the next step is to develop suitable directions and sample exercises for each different group of test exercises. Directions for tests must be clearly stated and in a vocabulary that the pupils can comprehend. If long difficult words have been used, the teacher should attempt to find synonyms which are in more common usage. In addition to the directions, it is important to provide sample exercises to give the pupils experience in employing the testing technique demanded. Pupils who have had little or no experience with objective exercises will be done an injustice unless they are given ample directions and practice on the types of exercises used. They may make low scores because they do not understand what to do, rather than

because they do not know the correct responses to the exercises. If many pupils fail to understand what to do, it is probable that the instructions are at fault. If this is not corrected the reliability of the test will certainly be lowered. The directions which accompany the samples of objective tests presented later in this chapter are examples of adequate statements.

79. Rating Exercises as to Difficulty.

After the questions have been developed and the directions and practice exercises perfected, the next step is to arrange the different groups of test exercises in the approximate order of difficulty from easiest to most difficult. This can be done roughly through inspection and rearrangement of the exercises by the teacher. If several teachers pool their judgments of the rankings from easiest to most difficult, the results will be more reliable. This arrangement of items in order of difficulty can be further refined after the test is given, by recording the number of pupils who respond correctly to the various items. Order of difficulty is quite important in a test because it saves the pupils' time and secures from them a better psychological reaction. The pupil is given an opportunity to answer first the exercises that are easier for him and he is not so likely to use all the testing time on difficult items and fail to answer many that he does know. The arrangement of items on the basis of difficulty probably increases the apparent reliability of the test.

80. Rearranging Items on the Basis of Difficulty.

The following true-false statements, taken from a longer test, are in the order in which they appeared when the test was first given. After the test was given and the pupils' responses were analyzed, a better order of arrangement in the test was possible. The numbers at the right indicate the order of increasing difficulty based on an analysis of the responses of 50 pupils. The exercise numbered "1" is the easiest item, i.e., was answered incorrectly by the smallest percentage of the class.

Original Order	Revised Order
1. Wipe moisture off of tools before putting them away.	6
2. The marking gauge is used to make a line parallel to an edge.	7
3. Sandpapering is done to get a smooth surface for finishing.	1
4. Stain may be applied with a cloth or a brush.	8
5. To produce a good surface for finishing, sandpaper across the grain.	9
6. Varnish is thinned with alcohol.	10
7. Good paint preserves the wood.	2

Original Order	Revised Order
8. Paint and varnish may be applied satisfactorily on damp surfaces.	3
9. Auger bits are graduated or numbered in sixteenths of an inch.	4
10. To bore a clean hole with an auger bit, bore through until the spur shows, and then finish boring from the other side.	5

It will be noted that only one of the easiest questions as indicated by an analysis of 50 pupil responses is in the first five items as listed in the original test arrangement.

81. Securing Reliability.

The next essential in developing an informal objective examination is to make it long enough to secure an acceptable reliability of measurement. Reliability is obtained by sampling over a wide range of content and by stating a large number of valid questions in objective forms which are within the mental and educational range of the pupils to be tested. Properly constructed objective tests of 75 to 100 or more exercises are usually highly reliable, whereas the ordinary six-, eight-, or ten-question essay-examination, with its limited sampling and subjective scoring, is almost never sufficiently reliable.

Sampling as a Factor in Reliability. The brevity of the statement and the ease with which the response is recorded make it possible for the student to respond to many more objective exercises in a specified period than to those of the discussion types. This makes it possible for the objective examination to cover a much wider area of subject-matter, or to cover a given area a great deal more intensively than is possible with the other type of exercise. The manner in which this factor of sampling operates to protect both the pupil and the teacher against the injustices of unreliable measurement is shown very clearly in Fig. 3, page 35, and is discussed in detail on pages 34 and 35.

Specific Hints on Securing Reliability in an Examination.—The following suggestions have been found useful in securing high reliability in objective informal tests:

1. Include from 50 to 100 items, each item being selected from definite units covering the entire area of the unit of the course. This step is closely related to securing high validity, but is considered here from the standpoint of reliability alone.

2. Make the questions objective in type. This eliminates the variable factor of the teacher's subjective judgment and gives assurance that all responses will be rated on the same basis.

3. Eliminate the dead weight from the test. Do not include items which are so easy that over 80 per cent of the class answer them correctly. Do not include items which are so difficult that less than

20 per cent of the class give the correct response. It is probable that test items which are missed by 80 per cent or more, or are missed by only 20 per cent or less of the class, do not differentiate pupil accomplishment adequately. These items can be determined by short tests during the term before they are put into the final test, or they can be eliminated after the test has been used once.

4. Control the conditions for giving the test. Define specific directions and conditions for administering the test.

5. Provide a key with the correct responses. It may be necessary to modify or give alternate answers on some completion exercises. The key, like other phases of the test, can be refined best after the test has been given.

82. Sample Objective Tests on Information Aspects of Elementary Woodwork.

A sample of the results of following through the steps in the construction of an objective examination as outlined in this chapter is shown in this section. The specimen is an experimental form of an objective test in woodworking which has been prepared and used by one of the authors in connection with his shop work. This test is in four parts requiring a total testing time of 42 minutes and having a possible total score of 94 points. Part I consists of 39 true-false items; Part II of four exercises in procedure-arrangement with a total point score of 22 points; Part III of 24 completion exercises; and Part IV of 9 multiple-response items. The reliability of this test based on 100 cases is .84. The total time requirements and the total possible score for each part are given in Table 26.

TABLE 26

CONTENT OF OBJECTIVE EXAMINATION

Part	Type of Exercise	Time Allowances	Possible Total Score
I	T-F	15	39
II	Pro.-Arr.	8	22
III	Compl.	12	24
IV	M-R	7	9
Total		42	99

Directions to Pupil

This test is divided into four parts. Specific directions are given for each part. The amount of time allowed is indicated below the directions. You are to stop working on each part when the teacher calls *time*. Do not begin the next part until the teacher reads the directions and gives the signal to begin work.

Do not waste time on a question you know nothing about; skip it, and go on to the next one. *Do not guess.*

You are now ready to study the directions for Part I. You are allowed *15 minutes* to complete Part I. Do not ask questions about the test after you begin work. If you break your pencil or need an eraser ask the teacher for one.

Part I. True-False

The following statements are to be answered by drawing a circle around the capital letter T or F which follows the statement. The letter T stands for *true*, and the letter F for *false*, or untrue.

If the statement is true, draw a circle around the T; if false, draw a circle around the F. The sample exercises are answered correctly.

Sample:	Nails are made of wood.	T	F
	Nails are made of metal.	T	F

You are allowed *15 minutes* for Part I. *Wait for the signal!*

Exercises

1. The surface of wood is planed to make it smooth.	T	F
2. Before putting away tools always wipe off any moisture that may be on them.	T	F
3. The marking gauge is used to make a line perpendicular to an edge.	T	F
4. Sandpapering on edges and surfaces is done in the direction of the grain.	T	F
5. Sandpapering is done to get a smooth surface suitable for finishing.	T	F
6. Sandpaper should be wrapped around a block when sanding flat surfaces.	T	F
7. Wood is stained in order to improve its appearance.	T	F
8. Stain may be applied with a cloth or a brush.	T	F
9. A well-made glue joint is weaker than any other part of the wood.	T	F
10. To obtain a good surface for finishing sandpaper across the grain.	T	F
11. No. 0 sandpaper is coarser than No. 00.	T	F
12. Varnish is thinned with alcohol.	T	F
13. A first-class job of finishing can be had with only one coat of varnish if it is put on thick enough.	T	F
14. Varnish can be smoothed down with fine steel wool.	T	F
15. In rubbing down varnish with powdered pumice stone, the pumice should be rubbed on dry.	T	F
16. A drawing board is made of oak, or some other hard wood.	T	F
17. The visible lines of an object are shown on a drawing by solid black lines.	T	F

18. The tee-square is used for making horizontal lines. T F
19. Thumbtacks should be hammered into place on the drawing board. T F
20. Templates are used to make, or mark out, a shape on a board. T F
21. The mortise of the mortise-and-tenon joint is the rectangular hole into which the tenon fits. T F
22. Good paint preserves the wood. T F
23. Paint and varnish may be applied satisfactorily on damp surfaces. T F
24. New wood soaks up much linseed oil. T F
25. New wood should have a priming coat applied before the finish paint is put on. T F
26. Paint and varnish are made from the same materials. T F
27. It is not necessary to brush paint out well when applying it. T F
28. If shellac is too thick, thin it out with turpentine. T F
29. Shellac is a slow-drying finish. T F
30. Shellac makes a waterproof finish. T F
31. It is easier to clean out a brush after varnishing with it if it is allowed to dry for 24 hours. T F
32. After paste wood filler has been applied, the surface must be rubbed across the grain. T F
33. Hand screws should be adjusted before any glue is applied to the pieces to be glued. T F
34. Handscrews hold best if the jaws are parallel to each other. T F
35. It is easier to drive in a screw if the screw-driver has a round tip, than if it has a square one. T F
36. Augur bits are graduated or numbered in thirty-seconds of an inch. T F
37. To bore a nice clean hole with an augur bit, bore through until the spur shows, and then finish boring from the other side. T F
38. A tee-bevel square is used when one wants to lay out an angle. T F
39. More accurate work can be done if knife lines are used, rather than pencil lines. T F

Stop and wait for the directions for Part II!

PART II. PROCEDURE-ARRANGEMENT

On this page are several jobs, and the steps necessary for doing the job. However, the steps are not placed in the correct order.

Decide which step should be done first, and place the number of the step in the first parenthesis, then the number of the second step in the second parenthesis, and so on until all the steps are down. The sample exercise is answered correctly.

Sample: To apply stain.

1. Let stand for 2-3 minutes.
2. Apply stain.
3. Smooth the surface with sandpaper.
4. Wipe off excess stain with cloth.
5. Select a suitable stain.

(3)....(5)....(2)....(1)....(4) correct order.

You are allowed *8 minutes* for Part II. *Wait for the signal!*

Exercises

1. List the following grades of sandpaper according to coarseness, placing the finest grade first.
   1. No. 0.
   2. No. ½.
   3. No. 2.
   4. No. 1½.
   5. No. 00.
   6. No. 1.

   ( )....( )....( )....( )....( )....( )

2. To square up a board.
   1. Plane a surface true; mark it No. 1.
   2. Plane one edge square with No. 1; mark it No. 2.
   3. Gauge and plane to thickness, square with edges and ends; mark No. 6.
   4. Cut to width and square other side with No. 1 and 3, and mark No. 5.
   5. Plane one end, and square with No. 1 and 2; mark No. 3.
   6. Cut to length, square other end with No. 1 and 2; mark No. 4.

   ( )....( )....( )....( )....( )....( )

3. Apply paint on new wood.
   1. Apply first coat of finish paint.
   2. Shellac the knots.
   3. Apply second coat of finish paint.
   4. Clean off any grease or dirt with cloth wet in benzine.
   5. Apply coat of priming paint.

   ( )....( )....( )....( )....( )

4. To bore a hole with a brace and bit.
   1. Fasten bit in brace.
   2. Withdraw bit and finish boring from opposite side.
   3. Mark location of hole.
   4. Bore through until spur shows on other side.
   5. Select proper size bit.

   ( )....( )....( )....( )....( )

Stop and wait for the directions for Part III!

Part III. Completion Exercises

Each of the following statements has one or two words left out. When the correct word or words are inserted in the blanks left for them, the sentences are specific and complete. The sample exercises are answered correctly.

Samples:

Nails are driven with a hammer.

Screws are driven with a ________. Here *screw-driver* is the correct word.

Think what word completes the sentence and write it in the blank space left for it.

You are allowed *12 minutes* for Part III.

Wait for the signal!

Exercises

1. A fine ________ is used for whetting a plane blade.
2. Sharpening chisels and plane irons on an oilstone removes the ________ ________.
3. The thickness of a shaving is regulated by the ________ ________.
4. The ________ holds the plane blade in place.
5. When plane is not in use lay it on its ________.
6. In starting a shaving cut with a plane, press ________ upon the knob of the plane.
7. In driving nails, at first use ________ steady strokes.
8. Inserting a wood block under the hammer when pulling nails prevents ________ the wood.
9. The ________ should be used to guide the saw when starting a cut.
10. Saws work easier when rubbed with ________ occasionally.
11. An augur bit is inserted into the ________ of the brace.
12. The number on the tang or shank of a bit indicates its ________.
13. A ________ ________ is used when boring a number of holes the same depth.
14. The teeth of a coping saw blade should point ________ handle.
15. The cross-cut saw is used for cutting ________ the grain.
16. The tool used for setting nails below the surface is called a ________ ________.
17. The cutting action of a ripsaw is like that of a number of ________.
18. The cutting action of a cross-cut saw is like a number of ________.
19. The ripsaw is used for cutting ________ the grain.
20. A ________ cornered file is used to sharpen saws.
21. Damp spongy wood requires a saw with plenty of ________.
22. The ________ works as a crank and holds the bit when boring.
23. The size of an augur bit in ________ of an inch may be found on the tang or shank.
24. The cut made by a saw is called a ________.

Stop and wait for the directions for Part IV!

Part IV. Multiple-Choice Exercises

Each of the statements below is answered correctly by one of the words following the sentence.

Determine which of the choice of words correctly answers the statement, and write the number of that word on the line at the end of the exercise. The sample exercise is answered correctly.

Sample: Sandpaper is made up of paper, glue, and ________

(1) brick dust, (2) sand, (3) emery, (4) gravel. 2

Sand is the correct answer, and the number of the word (2) is written on the line.

You are allowed 7 *minutes* for Part IV.

Wait for the signal!

Exercises

1. Shellac is thinned with ______
   (1) benzine, (2) turpentine, (3) water, (4) alcohol. ___
2. Oil stain is mixed with ______
   (1) alcohol, (2) turpentine, (3) water, (4) linseed oil. ___
3. Red cedar is the best wood for ______
   (1) dressers, (2) lamps, (3) chests, (4) bookracks. ___
4. Joinery is used in ______
   (1) plastering, (2) plumbing, (3) cabinet making, (4) bricklaying. ___
5. A good liquid to rub on tools to prevent rusting is ______
   (1) kerosene, (2) water, (3) machine oil, (4) turpentine. ___
6. A working drawing of an object shows the ______
   (1) corner view, (2) top view, (3) rear view, (4) bottom view. ___
7. Brushes that have been used in shellac should be cleaned in ______
   (1) oil, (2) turpentine, (3) alcohol, (4) gasoline. ___
8. Varnish brushes should be cleaned in ______
   (1) linseed oil, (2) shellac, (3) turpentine, (4) water. ___
9. Glued joints are commonly strengthened with ______
   (1) dowels, (2) rivets, (3) wire, (4) brads. ___

End of the test

The foregoing objective examination is designed to function as a paper-and-pencil test for measuring informational aspects of instruction in woodworking. Tests of this type can be developed by the industrial education teacher who will follow the principles outlined in this volume. The objectives of the course of study must be definitely identified. The rest of the process is largely the mechanical formulation of the selected items in suitable objective form. Such tests of information in industrial education are valuable as partial measures of achievement and teaching success, but alone, they are insufficient. They should be supplemented by performance tests.

II. CONSTRUCTION OF OBJECTIVE PERFORMANCE TESTS

The testing of performance is not new to industrial arts and industrial education. Manipulative trade tests were devised during the war and have been used with varying degrees of satisfaction in industry. The reliability of many of the early manipulative tests was low, and efforts to measure manipulative skill have not been as successful as the measurement of information by the use of the objective pencil-and-paper tests. A part of this difficulty has arisen from trying to apply pencil-and-paper techniques of test construction to manipulative-test construction without the necessary modifications in the administrative procedure.

83. Steps in Preparing Performance Tests.

In evaluating the available performance tests and in constructing performance tests in the general shop the authors have found the following steps worthy of careful consideration:

1. Analyze the course of study to determine exactly what qualities may be tested.
2. Decide what tools and materials will be necessary.
3. Prepare a number of test exercises or make a composite exercise that will offer the pupil an opportunity to provide an adequate sample of his work with each tool or instrument and type of material which it is desired to test.
4. Make a statement of procedure which tells the pupil exactly what to do in a vocabulary which is comprehensible at his grade level.
5. Prepare a set of general directions for the pupil before the test is administered.
6. Prepare directions for the examiner.
7. Devise a method of scoring the test which provides an adequate measure of the results of each tool or instrument.
8. Try out the test on a few students, and make the more obvious corrections.
9. Make two or more forms of the test.
10. Try out the test, and compute the reliability coefficient, standard deviation, probable error, etc.

For the purpose of illustrating the application of the principles of performance-test construction let us consider the measurement of the results of the following tool operations from a beginning woodworking course.

1. Planing: side, end grain.
2. Sawing: ripping, cross-cutting to a line.
3. Boring: perpendicular to a surface.
4. Squaring: a line around a block.
5. Measuring: to ⅛ inch with try-square and rule.
6. Gauge a line parallel to a surface.

It should be kept clearly in mind that this is not a test of technique or speed but a test of quality or accuracy. The question is, how accurately can a pupil modify materials with these tools regardless of the method of handling them or the time required.

The following tools are required: jack plane, try-square 6 inches, pencil, 24-inch folding rule, back saw, ripsaw, brace, ¼-inch bit,

bench with vise, bench hook, and marking gauge. The tools must be in first-class condition.

The following materials are required: first-quality white pine free of knots, 1 inch thick, surfaced on two sides, and ends sawed at an angle. It is most desirable to have the materials used in a test of this type of uniform quality and the pieces of stock used by the pupils of about the same size and shape.

A composite form of test exercise was selected for this test. Fig. 16 shows the working drawing for Form A, and Fig. 17 shows the working drawing for Form B. It will be observed that the main differences between the test exercises are in the dimensions; otherwise, there is the same opportunity for modifying wood with common tools.

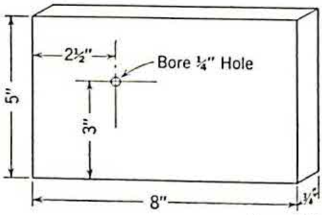


Fig. 16.

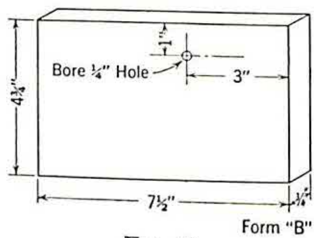


Fig. 17.

The validity of a performance test depends on providing samples of the pupil's work and enough of the sample to give an adequate measure of accuracy. Composite performance exercises are easier for the pupil to visualize if they represent some familiar object or toy, but it is seldom possible to do this without throwing the sampling of the various tool operations out of balance.

Woodworking Performance Test of Accuracy

Forms A and B

Directions to Examiner: It is essential that the pupil shall understand the exact procedure, and that he be able to visualize how the block is to look when finished. The following directions are recommended:

1. Read aloud and distinctly the directions to the pupil while the class follows silently. Answer any questions about the directions at this point.

2. Show the pupils a completed test block, and if they care to, let them examine it.

3. When there are no further questions, say, "Get ready. Hold up the test block. Begin work."

4. During the examination answer any questions about the steps in the procedure by rereading the step in question with the pupil.

5. Observe the pupils as they work to make certain that they are doing all the steps in the correct order.

6. Make certain that the proper tool is used where indicated, but do not tell the pupil how to use the tool.

7. Help any pupil having difficulty in interpreting the working drawing, but do not make any measurements on the test block for the pupil. This test measures ability to measure to 1/16 in. with a ruler, but is not a measure of ability to read drawings.

8. Take in the test block when the pupil has finished. The time is not important, for this is a test of quality or accuracy as it applies to modifying wood with simple hand tools.

Directions to Pupil. This is a test to determine how accurately you can use woodworking tools. The wood and all necessary tools will be given to you. The surfaces of the block of wood are numbered 1, 2, 3, 4, 5, 6. You will be given specific directions for doing the job and a working drawing that gives all the necessary dimensions. Do this project as accurately as you can. Do not waste time, but do not work too fast to do your best work. The steps must be done in the order given. After you begin work do not ask unnecessary questions, but if you are in doubt about a step in the procedure or a dimension on the working drawing ask the examiner. Write your *name* and *grade* in school on surface *No. 6* of the test block. *Do not begin work until the examiner gives the signal.*

Woodworking Performance Test—Form A

Part I

Procedure:[3]

1. Select face No. 1. Plane it square and true and to the thickness indicated on the working drawing (Fig. 16). When finished re-mark No. 1.

2. Select side No. 2. Plane it square and true to surface No. 1. When finished re-mark No. 2.

3. Select end No. 3. Plane it square and true to No. 1 and 2. When finished re-mark No. 3.

4. Measure from end No. 3 toward end No. 5, and square a sharp pencil line across the block to the length indicated in the working drawing. Saw off the waste material with a back saw so that the stock will be as nearly the required length as you can make it. *Do not plane.* Re-mark end No. 5.

5. From edge No. 2 gauge a line the length of the block, allowing the exact width as indicated on the working drawing. Rip as near the exact width as possible. *Do not plane.*

6. On surface No. 1 lay out the center for the hole and bore.

7. When you have finished take your block to the examiner.

84. Scoring Performance Tests.

A try-square, a 1/4-inch dowel 3 inches long, and a scale measuring in sixty-fourths are the tools used in scoring. The test is scored in units of sixty-fourths of an inch. If a measurement is more than 10/64 off, it is given a score of zero. If it is 1/64 off, it is given 9 points, 5/64, 5 points, etc., as shown in Table 27.

[3] Procedures for Forms A and B are identical although details are different.

This test has been used a number of times by the authors in shop measurement. The correlation of Form A with Form B on 30 cases shows a reliability coefficient as high as .90. As has been pointed out, this test is not concerned with speed and technique, but only with the accuracy or quality of work. It is not difficult to construct valid and reliable manipulative tests if the test worker keeps clearly in mind the different measurable factors in industrial education. However, it is well to remember that it requires skill to administer a manipulative test. The examiner must follow the technique carefully. A fine tool in the hands of the unskilled worker will not necessarily produce an acceptable result.

TABLE 27

Limits of Tolerance	Point Scores
0/64	10
1/64	9
2/64	8
3/64	7
4/64	6
5/64	5
6/64	4
7/64	3
8/64	2
9/64	1

SUMMARY

The outline of the curricular content of the unit affords the basis for validating the content of the test. With these objectives and the details of the informational content of the course as a background, a paper-and-pencil test is easily formulated by using the types of testing techniques discussed in Chapter X.

Knowledge of the informational aspects of courses in industrial education is only one of the important outcomes. Actual production in the shop or laboratory is quite another important outcome, which, within certain limits, is measurable. Quality in shop work may be rated by inspection, by actual measurement of dimensions, and by judgments based on quality scales. Inspection, if based on personal judgments unsupported by objective criteria of quality, is highly inaccurate. Actual measurement of the physical qualities (dimensions, etc.) is probably the most objective. However, there are certain other qualities in shop products which are not mere matters of dimensions or accuracy of measurement or tool work. The evaluation of such qualities require the use of a rating scale. Such quality scales are not merely useful for the measurement of accomplishment, but they are particularly helpful in the development of an appreciation of quality on the part of the pupil.

SUMMARY EXERCISES FOR DISCUSSION

1. List and illustrate the constructive steps in validating an objective information test in an industrial subject.
2. Following the general plan presented in the section in this chapter entitled

Course of Study Outline in Woodwork, prepare a similar list of specific objectives for some other phase of industrial education.

3. What new suggestions can you add to the list of specific hints on securing reliability in an examination?
4. Prepare at least ten objective test items of each of the four types illustrated on pages 140 to 144 of this chapter, using any industrial education subject-matter except woodworking.
5. Prepare a performance test in metal working, printing, or mechanical drawing following the steps outlined in this chapter.

SELECTED REFERENCES

GARRETT, HENRY E., *Statistics in Psychology and Education.* New York: Longmans, Green and Company, 1926.

KELLEY, TRUMAN L., *Statistical Method.* New York: The Macmillan Company, 1923.

ODELL, C. W., *Traditional Examinations and New-Type Tests.* New York: The Century Company, 1928.

PATERSON, ELLIOTT, ANDERSON, *et al., Minnesota Mechanical Ability Tests.* The University of Minnesota Press, 1930, pp. 188-202.

RUCH, G. M., *The Objective, or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929.

RUCH, G. M., and RICE, G. A., *Specimen Objective Examinations.* Chicago: Scott, Foresman and Company, 1930.

SYMPOSIUM ON TESTING, *Epsilon Pi Tau Review,* Vol. II. The Ohio State University, Columbus, 1931.

CHAPTER XII

CONSTRUCTION AND USE OF SCALES FOR RATING INDUSTRIAL EDUCATION PROJECTS

I. PROJECT RATING SCALES

85. Need for Scales for Rating Industrial Education Projects.

The rating of shop projects and drawings is a difficult measurement problem which is faced by practically all industrial education teachers. The assignment of an objective rating to a shop product such as a chair, a radio, a table, a lamp, a funnel, a dustpan, a cement bird-bath, or an isometric drawing calls for the keenest of discrimination and for some method of objectifying standards of quality. Such products of the shop are composed of many different parts, which, combined, reflect quality in the object. For example, in the funnel, there are such factors as the forming, turning, wiring, seaming, and soldering. All these operations must be well done if the funnel is to have quality. Moreover, it must be made of the proper size and of suitable material. Thus size, shape, quality of material, suitability of material, and quality of workmanship as revealed in many small details must be recognized and evaluated by the judge. Each additional characteristic sets up a number of possible variations in quality.

The difficulty of grading or rating such products appears to be related to the degree to which they vary from the typical. For example, shop and drawing projects which are well executed and those which are very inaccurate present a simpler marking problem than those which are made up of mixtures of good, bad, and indifferent qualities. This may be made clear by a concrete illustration. Let us assume that a number of blocks or dominoes are to be arranged in a straight line one inch apart with all faces parallel. If all the units are arranged correctly it is not difficult to see that this is true. In a like manner, if a shop project is very well done, it tends to radiate perfection. Likewise, a very poor project is not difficult to distinguish. However, the most difficult problem of measurement presents itself when part of the dominoes are set up correctly, a few of them are down, and all the rest are in varying degrees of alignment. The shop project which shows excellent design, a poor finish, square edges, weak joints, rough

surfaces, and carefully rounded corners is likewise more difficult to judge. If the design is given undue consideration, the project will be rated too high. If only the joints and surfaces are considered, the pupil may fail on the project.

86. Constructing a Project Rating Scale.

The chief problem confronting shop and drawing teachers in their rating of projects is the definition of objective standards of quality. Industrial arts teachers have recognized the need for better methods of rating shop projects. Many realize that the rating of shop and drawing problems is so subjective and unreliable that it is difficult to assign the proper rating to a pupil's project. In general, the suggestions for improvement in the reliability of rating shop projects indicate the desirability of combining the judgments on the different parts of the projects into a complete rating, and having the projects rated by three or more qualified judges.

The authors have found the following principles helpful in constructing a rating scale for shop and drawing projects:

1. Make a careful analysis of the course of study for the purpose of selecting the factors to put in the rating scale. In general, the items rated are the changes made in materials by the use of tools or instruments, fasteners, and finishes.

Example (from eighth-grade general woodworking unit): Utility, design, proportion, nailing, squareness, dimensions, screws, joints, glued joints, boring, sawed edges, planed edges, sanding, and finish.

Example (from ninth-grade drawing): Neatness, dimensions, arrowheads, lines, accuracy, French curves, placement, joints, lettering, and circles.

2. Group the factors into classes according to method of rating to be used.

Example: Woodwork.

- Inspection.
  - Utility.
  - Design.
  - Proportion.
  - Finish.
- Physical measurement.
  - Squareness.
  - Dimensions.
- Rating scale or inspection.
  - Nailing.
  - Screw joints.
  - Glue joints.

Boring.
Sawed edges.
Planed edges.
Sanding.
Boring.
Wood filing.

Example: Drawing.

- Inspection.
  - Neatness.
  - Placement.
  - Arrowheads.
- Physical measurement.
  - Circle.
  - Accuracy.
  - Dimensions.
- Rating scale or inspection.
  - Lettering.
  - Lines.
  - Numbering.

3. Put the factors into a rating scale so that each part of the project can be given an individual objective rating and the ratings combined.

Inspection by a critical and observant judge may reveal many defects in quality. Checks on the physical measurements of products by means of the rule and the square are very objective. Pieces of material can be tested for squareness, thickness, width, and length. There are, however, numerous quality factors which are less tangible and are rated better with quality scales especially developed for the purpose. Splicing, lettering, soldering, and sawing are examples of such qualities.

4. Prepare a set of directions for using the project rating scale.

The project rating scale needs a carefully prepared set of directions which explain in simple, direct language just how the scale is used. This should include the method of recording the ratings on the individual parts of the rating scale, the tools, and the quality scales needed, as well as a statement of the method of arriving at the complete score. A place should also be provided for recording the necessary information about the pupil and the project, as, for example, the student's name, grade in school, date of finishing project, name of project, name or judge, and total score.

5. Prepare a key for transforming the distance ratings into objective values for use in computing the composite rating. This step is not necessary when the scale units are numbered on the scale.

87. Typical Project Rating Scales.

On the following pages are excerpts from a project rating scale for mechanical drawing which the authors have found helpful in their classes and which may serve to illustrate the application of the principles discussed in this chapter.[1]

I

RATING SCALE FOR MECHANICAL DRAWINGS

Pupil's name ____________________

Drawing ____________________

School ____________________ Date __________

Name of judge ____________________ Score __________

Directions: The information required to rate the items is obtained by inspection, quality scales, and physical measurement. Each item is rated on the basis of 10 points. The total rating of the drawing is the sum of all the item ratings which apply to the drawing.

Mark the items in the order in which they appear.

Draw a circle around the figure on the scale to indicate your rating of the item. The scale is divided into ten (10) equal parts. The right side (10) indicates the highest mark; the center, average; the left, the lowest. A profile of the ratings may be made if desired by connecting the numbers representing the ratings assigned to each item which applies to the specific project.

Example: If you think *Utility* under *I The Drawing* should get a mark of *8, then draw a circle around the figure 8, thus:*

1. The Drawing.
   1. Utility __________ 1 2 3 4 5 6 7 8 9 10

Instruments: Architect's scale, compass.

Quality Scales: The judges should have quality scales for lettering figures and lines each with ten or more samples of known value.[2]

I. The layout.

1. The placing on paper __________ 1 2 3 4 5 6 7 8 9 10
   Is the object placed in such a manner as to permit a clear-cut drawing?
2. Geometric construction __________ 1 2 3 4 5 6 7 8 9 10
   How good is the instrument work?
3. Trueness in meeting lines __________ 1 2 3 4 5 6 7 8 9 10
   Do the lines meet truly and completely?
4. Construction lines
   a. Accuracy __________ 1 2 3 4 5 6 7 8 9 10
   By how many thirty-seconds do the lines miss?
   b. Quality __________ 1 2 3 4 5 6 7 8 9 10
   Are the lines clean and sharp cut?

[1] The latter part of this chapter is devoted to a discussion of a reliable method of constructing quality scales.

[2] Inspection must be used when quality scales are not available.

II. The finished drawing (pencil or ink).

1. Circles.

a. Neatness ______ 1 2 3 4 5 6 7 8 9 10

Are all lines clean-cut and uniform?

b. Trueness ______ 1 2 3 4 5 6 7 8 9 10

Are the circles truly drawn?

c. Accuracy ______ 1 2 3 4 5 6 7 8 9 10

Does the diameter vary from the true dimension more than 1/32 inch?

d. Tangency ______ 1 2 3 4 5 6 7 8 9 10

Do tangent lines fit in perfectly?

2. Arcs and curves.

a. Neatness ______ 1 2 3 4 5 6 7 8 9 10

Are the lines clean-cut and uniform?

b. Trueness ______ 1 2 3 4 5 6 7 8 9 10

Are the curves and arcs truly drawn?

c. Accuracy ______ 1 2 3 4 5 6 7 8 9 10

Do lines vary more than 1/32 inch from given dimension?

d. Tangency ______ 1 2 3 4 5 6 7 8 9 10

Do tangent lines fit in perfectly?

e. Completeness ______ 1 2 3 4 5 6 7 8 9 10

Do the lines run up completely?

3. Horizontal lines.

a. Neatness ______ 1 2 3 4 5 6 7 8 9 10

Are the lines clean-cut and uniform?

b. Accuracy ______ 1 2 3 4 5 6 7 8 9 10

Do lines vary more than 1/32 inch from given dimension?

c. Trueness ______ 1 2 3 4 5 6 7 8 9 10

Are all lines straight and horizontal?

d. Completeness ______ 1 2 3 4 5 6 7 8 9 10

Do the lines run short?

4. Vertical lines.

a. Neatness ______ 1 2 3 4 5 6 7 8 9 10

Are all lines clean-cut and uniform?

b. Accuracy ______ 1 2 3 4 5 6 7 8 9 10

Do lines vary more than 1/32 inch from given dimension?

c. Trueness ______ 1 2 3 4 5 6 7 8 9 10

Are all lines straight and vertical?

d. Completeness ______ 1 2 3 4 5 6 7 8 9 10

Do lines run up completely?

5. Skew lines.

a. Neatness ______ 1 2 3 4 5 6 7 8 9 10

Are the lines clean-cut and uniform?

b. Accuracy ______ 1 2 3 4 5 6 7 8 9 10

Do the dimensions vary more than 1/32 inch from given dimensions?

c. Trueness ______ 1 2 3 4 5 6 7 8 9 10

Are the lines straight and to the points?

d. Completeness 1 2 3 4 5 6 7 8 9 10
Do lines run up completely?

6. Dimension lines.
a. Placing 1 2 3 4 5 6 7 8 9 10
Are they placed correctly and conspicuously?
b. Quantity 1 2 3 4 5 6 7 8 9 10
Are the necessary lines in?
c. Quality 1 2 3 4 5 6 7 8 9 10
Are they the true lines as to accuracy?
d. Spacing 1 2 3 4 5 6 7 8 9 10
Are they spaced too close or too much?
e. Correctness 1 2 3 4 5 6 7 8 9 10
Are they correctly put in?
f. Arrowheads.
(1) Quality 1 2 3 4 5 6 7 8 9 10
Are they neat and trim?
(2) Accuracy 1 2 3 4 5 6 7 8 9 10
Do they come up to the extension lines?
g. Extension lines 1 2 3 4 5 6 7 8 9 10
Do they run into the object?

7. Dimensions.
a. Legibility 1 2 3 4 5 6 7 8 9 10
Can they be read?
b. Correctness 1 2 3 4 5 6 7 8 9 10
Are they correctly put in?

8. Notes.
a. Legibility 1 2 3 4 5 6 7 8 9 10
Can they be read?
b. Clearness 1 2 3 4 5 6 7 8 9 10
Do they state exactly what is wanted?
c. Lettering
(1) Uniformity 1 2 3 4 5 6 7 8 9 10
Are the letters uniform as to height and slant?
(2) Appearance 1 2 3 4 5 6 7 8 9 10
How are the letters formed?
(3) Firmness 1 2 3 4 5 6 7 8 9 10
Are the strokes firm?

III. Summing up the drawing.
1. Utility 1 2 3 4 5 6 7 8 9 10
Can the drawing be used?
2. Value 1 2 3 4 5 6 7 8 9 10
Is the drawing of any aid in making the project drawn?
3. Appearance 1 2 3 4 5 6 7 8 9 10
Is the drawing executed in a neat, professional manner?
4. Completeness 1 2 3 4 5 6 7 8 9 10
Is the drawing complete?

88. Using Project Rating Scales.

Not only does the project rating scale afford a more objective means of rating projects from the shop and drawing-room, but it is itself a valuable teaching device. It aids the pupil in developing a proper appreciation of quality in workmanship. Its use by students in the rating of their own projects and those of their classmates gives them valuable opportunities for developing habits of careful analysis and experience in judging quality in workmanship.

The best teaching results are obtained with the project rating scale when it is used during the time the project is being made. The items in the project scale should be arranged in an order which facilitates the use of the project rating scale parallel with the development of the project itself. The finish, utility, design, and proportion of a project are better judged after the project is completed. The results of sawing, gluing, squaring, screwing, turning, forming, soldering, riveting, splicing, etc., are easier to judge just after they are completed and before they have been obscured or modified by other tool operations or parts of the project. Sawed or planed edges are often modified by filing, scraping, and sanding. The quality of these operations must be rated at a time when they give a true picture of the pupil's proficiency.

The general quality of a project is determined by the sum total of the operations which go into its making. It is well for a pupil to realize this, and to use the diagnostic value of the rating scales to check the results of the tool operations. If a pupil is unable to set a rivet, make a splice, or bore a hole in an acceptable manner, both the pupil and the instructor should be aware of the fact. The pupil may need remedial work, and even though he may never be able to achieve a high standard of workmanship he can and should develop an appreciation of what constitutes quality in workmanship. One generally accepted method of developing an appreciation of quality is to allow the pupil to modify materials and to judge the results of his own efforts and those of others in relation to acceptable standards.

Teachers of industrial education need training and practice in the construction and use of project rating scales. This is a function that should be taken over by the teacher-training institutions and supervisory officers. In the event that the proper training is not available, the progressive industrial education teacher should make project rating scales which are based on his course of study and practice their use. In any case a carefully prepared project rating scale reduces the variation of marks and provides a valuable teaching device which cannot be ignored by those who wish to do superior work in their shops and drafting-rooms. A carefully prepared project rat-

ing scale is more objective than the usual mark assigned by the teacher, but it is distinctly more meaningful when the ratings of from five to ten judges are pooled. When the pooled judgments of such a number of trained observers are used, reliability coefficients of .90 or better are obtained. This is a satisfactory reliability of measurement and produces a rating comparable to the best objective ratings in other fields of instruction.

II. QUALITY SCALES FOR SHOP PRODUCTS

89. Constructing Quality Scales.

Quality scales are useful measurement and teaching instruments which the shop teacher can easily develop and use in his course. Such scales are a very useful means of evaluating certain types of work done in the class. If made available for inspection and comparison, they serve to set up standards for the students themselves to attain, thus aiding the pupil in developing an appreciation of quality. The pupils may rate their own and other pupils' work, thus gaining real experience in rating and further developing a concept of what constitutes real quality in workmanship.

The teacher of woodworking will find that quality scales dealing with such skills as sawing, boring-exit edge, fastening with screws, gluing, planing, sanding, and nailing are very helpful instruments to use in the more exact evaluation of shop products. The specimens constituting the scales can be mounted on suitable panels to be conveniently available for inspection and use. The teacher of sheet-metal work needs quality scales showing the range of workmanship in riveting, soldered lap-seam work, wiring, locked-seam work, and turning. This list may be reduced or expanded depending on the type of course being taught.

Quality scales have been widely used in rating such school products as writing, lettering, and drawing. Scales of this type can be reproduced and widely distributed, but quality scales dealing with pieces of material are not so easily duplicated. Quality scales for shop use should be made up of actual specimens of the products themselves. They should be available to the students so that the specimens may be seen and handled. Such scales may be photographed, but in use the picture lacks the satisfaction resulting from a scale composed of actual samples of the work to be rated.

Because of the real value which quality scales have in the measurement of industrial education and in the development of the pupil's

concept of quality, a method for developing such scales is presented in this chapter.

90. Steps in Making a Quality Rating Scale.

The following steps are essential in developing a quality scale for shop products:

1. Secure samples of the type of product to be used in the scale.
2. Arrange the samples in order of merit as determined by the judgment of ten or more competent judges.
3. Determine the percentages of judges rating a given sample as *better than* each other sample.
4. Arrange the specimens in order from best to poorest on the basis of these composite judgments.
5. Find the deviation of the percentages from the median (50 per cent), retaining the sign of the deviation.
6. Calculate the scale differences between the successive samples.
7. Assume a zero point, and place all samples along the linear scale of values.
8. Select eight to twelve of the samples which nearest approach uniform differences in quality for the quality scale, assign the proper quality values to them, and mount them for use in the shop.

Each of these essential steps in constructing a quality scale is discussed and illustrated in succeeding sections of this chapter. A quality scale showing degrees of merit in soldering a lap seam is used for illustrative purposes. Fig. 19 shows the values assigned in the completed scale.

91. Securing Samples.

Twenty to forty representative samples ranging widely in quality will usually be adequate for the construction of a satisfactory quality rating scale for use in an industrial education course. In securing samples for such a scale, it is essential that some of the samples be superior, some poor, and some average in quality. Samples of average quality are easy to secure, but it may be more difficult to get a proper number of excellent and poor ones. If twenty to forty samples have been secured and, after a superficial inspection, it is obvious that there is a shortage of poor and excellent ones, the instructor should make a few samples which are excellent and a few which are very poor and add them to the list. This practice is defensible, as it might be necessary to sample very widely in order to secure enough specimens showing the required quality range. The use of many more than forty samples makes it more difficult for the judges to rate them carefully

and also adds considerably to the amount of calculation necessary in the derivation of the scale.

Twenty-two samples were used in construction of the scale in Fig. 19. It shows the quality of soldering on a lap seam. No samples were added to this group. The students who made the samples varied widely in their ability and background.

92. Securing Independent Estimates of Relative Quality of the Samples.

The samples to be used in making the scale should be lettered or numbered to make them readily identifiable. Names of pupils who made them should not be attached, since they might influence the judges' rating. After the samples are properly labeled, they should be given to the judges individually in random order. The judges are instructed to arrange them from poorest to best by comparison. The judges may be shop teachers or tradesmen, but they should be persons competent to rate the quality of the samples. The number of judges should not be less than nine or ten. A larger number would be better. If ten (or multiples of ten) judges can be secured, the calculation of percentages is simplified. It might be permissible for an isolated teacher who is unable to secure the cooperation of that number of qualified judges to have the same judge ratė the samples more than once. If this procedure is followed, it is advisable to allow a day or two between ratings to reduce the influence of memory in placing the samples.

Table 28 gives the results obtained when nine judges ranked the twenty-two soldered lap joints used in making the scale illustrated here. Each of the twenty-two samples is designated by a letter.

TABLE 28

RANKING BY NINE JUDGES

	High 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Low 20	21	22
Judge 1	I	P	U	R	B	A	G	N	E	S	L	C	Q	V	O	J	F	D	K	M	H	T
Judge 2	B	I	S	R	U	P	G	Q	A	L	E	N	C	J	M	V	K	O	H	D	F	T
Judge 3	P	B	A	N	I	U	S	J	Q	L	V	E	R	G	K	C	O	M	H	F	D	T
Judge 4	B	P	Q	E	I	R	U	G	C	S	N	A	L	J	H	M	O	D	V	K	T	F
Judge 5	B	R	P	G	Q	S	U	I	A	J	E	V	L	K	M	O	F	H	N	D	C	T
Judge 6	B	P	Q	I	R	A	G	E	U	S	V	J	L	N	O	K	C	M	F	D	H	T
Judge 7	B	I	R	C	A	P	U	G	S	H	J	Q	E	V	L	M	K	F	N	O	D	T
Judge 8	B	I	Q	S	U	P	R	J	C	G	A	N	V	M	O	E	K	F	L	D	H	T
Judge 9	B	I	J	C	A	M	U	P	R	G	S	Q	V	L	N	E	K	O	F	D	T	H

93. Determining Percentage Ratings of Judges.

The ratings of the twenty-two specimens by the nine judges are tabulated in Table 28. The next step is the determination of the percentage of judges rating a given sample as *better than* every other sample. Table 29 gives these data for the nine judges and twenty-two samples used in developing the quality scale for soldering lap joints. It will be noted that the samples are rated in alphabetical order. This table is to be read as follows: A is better than A no times because they are the same sample, but B is rated better than A nine times which means all the judges considered it better than A, etc. It will be noted that C is rated better than A four times, better than B zero times, etc. This table, showing the number of judges rating each sample in relation to every other sample, is the basis for the next step in constructing the scale.

The next step is to change the ratings to percentages. Table 30 shows the ratings in Table 29 after they have been changed to percentages. Specimen B is rated better than A by nine judges, or 100 per cent. Specimen C is rated as better than A by four judges, or 44.4 per cent. This procedure when completed gives the results as reported in Table 30.

94. Determining Order of Merit of Samples.

The rank order of the samples is determined by referring to the sum of the ratings for each sample in Table 29. The sample with the highest total rating is highest in quality, the one with the next highest is second, and so on. This gives useful information for construction of a quality scale, since it indicates the relative order as given by the combined judgment of nine judges. However, it does not tell the exact distance between each sample along a known scale. It is necessary to know the relation of the respective samples to each other before a number of samples can be selected to represent approximately equal distances on the scale.

95. Determining the Scale Differences.

The first step is to find the deviation of the percentages from the median (50 per cent). This is accomplished by subtracting from the median (50 per cent) the percentage values for each rating as given in Table 30. These percentages are then expressed as positive or negative deviations from the median (Table 31). All cases of 100 per cent and zero per cent are omitted from the table since they do not operate to affect the results. After the technique of developing quality scales is mastered, the student will find it convenient to omit the preparation of Table 31.

TABLE 29

RATING OF SPECIMENS BY JUDGES

(Table shows the number of judges rating specimens indicated in the horizontal lists as better than those in the vertical columns.)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
A		9	4	0	1	0	4	0	8	2	0	0	0	1	0	7	5	7	4	0	5	0
B	0		0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	1	0	0	1	0
C	5	9		1	6	1	6	1	9	5	3	5	1	5	2	7	6	8	6	0	7	3
D	9	9	8		9	7	9	5	9	9	7	9	8	9	9	9	9	9	9	0	9	8
E	8	9	3	0		0	7	1	8	5	0	3	2	4	1	9	8	7	6	0	7	3
F	9	9	8	2	9		9	4	9	9	8	8	8	7	8	9	9	9	9	1	9	9
G	5	9	3	0	2	0		0	8	3	0	1	1	1	0	9	4	9	3	0	7	1
H	9	9	8	4	8	5	9		9	8	7	8	7	7	7	9	8	9	9	1	9	7
I	1	8	0	0	1	0	1	0		0	0	0	0	1	0	4	3	1	1	0	1	0
J	7	9	4	0	4	0	6	1	9		0	3	0	4	1	8	6	7	8	0	8	2
K	9	9	6	2	9	1	9	2	9	9		8	5	7	4	9	9	9	9	0	9	9
L	9	9	4	0	6	1	8	1	9	6	1		2	4	1	8	9	8	9	0	9	5
M	9	9	8	1	7	1	8	2	9	9	4	7		6	3	8	8	8	8	0	8	6
N	8	9	4	0	5	2	8	2	8	5	2	5	3		1	9	7	8	7	0	8	4
O	9	9	7	0	8	1	9	2	9	8	5	8	6	8		9	9	9	9	0	0	8
P	7	7	2	0	0	0	0	0	5	1	0	1	1	0	0		1	3	2	0	3	0
Q	4	9	3	0	1	0	5	1	6	3	0	0	1	2	0	8		5	5	0	5	0
R	2	8	1	0	2	0	0	0	8	2	0	1	1	1	0	6	4		3	0	4	1
S	5	9	3	0	3	0	6	0	8	1	0	0	1	2	0	7	4	6		0	6	0
T	9	9	9	9	9	8	9	8	9	9	9	9	9	9	9	0	9	9	9		9	9
U	4	8	2	0	2	0	2	0	8	1	0	0	1	1	0	6	4	5	3	0		0
V	9	9	6	1	6	0	8	2	9	7	0	4	3	5	1	9	9	8	9	0	9	
Totals	137	194	93	20	98	27	123	31	165	102	37	80	60	84	47	152	131	145	128	2	133	76
	18	22	11	2	12	3	14	4	21	13	5	9	7	10	6	20	16	19	15	1	17	8

TABLE 30

PERCENTAGE OF JUDGES RATING EACH SPECIMEN BETTER THAN ANOTHER

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
A	X	100	44.4	0	11.1	0	44.4	0	88.9	22.2	0	0	0	11.1	0	77.8	56.5	77.8	44.4	0	55.5	0
B	0	X	0	0	0	0	0	0	11.1	0	0	0	0	0	0	22.2	0	11.1	0	0	11.1	0
C	55.6	100	X	11.1	66.7	11.1	66.7	11.1	100	55.5	33.3	55.5	11.1	55.5	22.2	77.8	66.7	88.9	66.7	0	77.8	33.3
D	100	100	88.9	X	100	77.8	100	55.6	100	100	77.8	100	88.9	100	100	100	100	100	100	0	100	88.9
E	88.9	100	33.3	0	X	0	77.8	11.1	88.9	55.6	0	33.3	22.2	44.4	11.1	100	88.9	77.8	66.7	0	77.8	33.3
F	100	100	88.9	22.2	100	X	100	44.4	100	100	88.9	88.9	88.9	77.8	88.9	100	100	100	100	11.1	100	100
G	55.5	100	33.3	0	22.2	0	X	0	88.9	33.3	0	11.1	11.1	11.1	0	100	44.4	100	33.3	0	77.8	11.1
H	100	100	88.9	44.4	88.9	55.5	100	X	100	88.9	77.8	88.9	77.8	77.8	77.8	100	88.9	100	100	11.1	100	77.8
I	11.1	88.9	0	0	11.1	0	11.1	0	X	0	0	0	0	11.1	0	44.4	33.3	11.1	11.1	0	11.1	0
J	77.8	100	44.4	0	44.4	0	66.7	11.1	100	X	0	33.3	0	44.4	11.1	88.9	66.7	77.8	88.9	0	88.9	22.2
K	100	100	66.7	22.2	100	11.1	100	22.2	100	100	X	88.9	55.5	77.8	44.4	100	100	100	100	0	100	100
L	100	100	44.4	0	66.7	11.1	88.9	11.1	100	66.7	11.1	X	22.2	44.4	11.1	88.9	100	88.9	100	0	100	55.6
M	100	100	88.9	11.1	77.8	11.1	88.9	22.2	100	100	44.4	77.8	X	66.7	33.3	88.9	88.9	88.9	88.9	0	88.9	66.7
N	88.9	100	44.4	0	55.5	22.2	88.9	22.2	88.9	55.5	22.2	55.5	33.3	X	11.1	100	77.8	88.9	77.8	0	88.9	44.4
O	100	100	77.8	0	88.9	11.1	100	22.2	100	88.9	55.5	88.9	66.7	88.9	X	100	100	100	100	0	0	88.9
P	77.8	77.8	22.2	0	0	0	0	0	55.5	11.1	0	11.1	11.1	0	0	X	11.1	33.3	22.2	0	33.3	0
Q	44.4	100	33.3	0	11.1	0	55.5	11.1	66.7	33.3	0	0	11.1	22.2	0	88.9	X	55.5	55.5	0	55.5	0
R	22.2	88.9	11.1	0	22.2	0	0	0	88.9	22.2	0	11.1	11.1	11.1	0	66.7	44.4	X	33.3	0	44.4	11.1
S	55.5	100	33.3	0	33.3	0	66.7	0	88.9	11.1	0	0	11.1	22.2	0	77.8	44.4	66.7	X	0	66.7	0
T	100	100	100	100	100	88.9	100	88.9	100	100	100	100	100	100	100	0	100	100	100	X	100	100
U	44.4	88.9	22.2	0	22.2	0	22.2	0	88.9	11.1	0	0	11.1	11.1	0	66.7	44.4	55.5	33.3	0	X	0
V	100	100	66.7	11.1	66.7	0	88.9	22.2	100	77.8	0	44.4	33.3	55.5	11.1	100	100	88.9	100	0	100	X

TABLE 31

PERCENTAGE DEVIATIONS FROM MEDIAN (50 PER CENT)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
A	X		−5.6		−38.9		−5.6		+38.9	−27.8				−38.9		+27.8	+5.6	+27.8	−5.6		+5.6	
B		X							−38.9							−27.8		−38.9			−38.9	
C	+5.6		X	−38.9	+16.7	−38.9	+26.7	−38.9		+5.6	−16.7	+5.6	−38.9	+5.6	−27.8	+27.8	+16.7	+38.9	+16.7		+27.8	−16.7
D			+38.9	X		+27.8		+5.6			+27.8		+38.9									+38.9
E	+38.9		−16.7		X		+27.8	−38.9	+38.9	+5.6		−17.6	−27.8	−5.6	−38.9		+38.9	+27.8	+16.7		+27.8	−16.7
F			+38.9	−27.8		X		−5.6			+38.9	+38.9	+38.9	+27.8	+38.9					−38.9		
G	+5.6		−16.7		−27.8		X		+38.9	−16.7		−38.9	−38.9	−38.9	·		−5.6		−16.7		+27.8	−38.9
H			+38.9	−5.6	+38.9	+5.6		X		+38.9	+27.8	+38.9	+27.8	+27.8	+27.8		+38.9			−38.9		+27.8
I	−38.9	+38.9			−38.9		−38.9		X					−38.9		−5.6	−16.7	−38.9	−38.9		−38.9	
J	+27.8		−5.6		−5.6		+16.7	−38.9		X		−16.7		−5.6	−38.9	+38.9	+16.7	+27.8	+38.9		+38.9	−27.8
K			+16.7	−27.8		−38.9		−27.8			X	+38.9	+5.6	+27.8	−5.6							
L			−5.6		+16.7	−38.9	+38.9	−38.9		+16.7	−38.9	X	−27.8	−5.6	−38.9	+38.9		+38.9				+5.6
M			+38.9	−38.9	+27.8	−38.9	+38.9	−27.8			−5.6	+27.8	X	+16.7	−16.7	+38.9	+38.9	+38.9	+38.9		+38.9	+16.7
N	+38.9		−5.6		+5.6	−27.8	+38.9	−27.8	+38.9	+5.6	−27.8	+5.6	−16.7	X	−38.9		+27.8	+38.9	+27.8		+38.9	−5.6
O			+27.8		+38.9	−38.9		−27.8		+38.9	+5.6	+38.9	+16.7	+38.9	X							+38.9
P	+27.8	+27.8	−27.8						+5.5	−38.9		−38.9	−38.9			X	−38.9	+16.7	−27.8		−16.7	
Q	−5.6		−16.7		−38.9		+5.6	−38.9	+16.7	−16.7			−38.9	−27.8		+38.9	X	+5.6	+5.6		+5.6	
R	−27.8	+38.9	−38.9		−27.8				+38.9	−27.8		−38.9	−38.9	−38.9		+16.7	−5.6	X	−16.7		−5.6	−38.9
S	+5.6		−16.7		−16.7		+16.7		+38.9	−38.9			−38.9	−27.8		+27.8	−5.6	+16.7	X		+16.7	
T						+38.9		+38.9												X		
U	−5.6	+38.9	−27.8		−27.8		−27.8		+38.9	−38.9			−38.9	−38.9		+16.7	−5.6	+5.6	−16.7		X	
V			+16.7	−38.9	+16.7		+38.9	−27.8		+27.8		−5.6	−16.7	+5.6	−38.9			+38.9				X

The percentage values which are given in Table 31 are next referred to any table showing the fractional parts of the total area of the normal probability curve corresponding to designated distances on the base line. Such a table as Table X on page 91 of Garrett's *Statistics in Psychology and Education* (Longmans, Green and Company, 1927) or Table 51 on page 219 of Thorndike's *Mental and Social Measurements* (Teachers College, Columbia University, 1916) affords the necessary data. These tables show, for example, that Specimen C, which was rated as better than Sample A by 44.4 per cent of the judges (Table 30) lies below the quality of Sample A a distance of 5.6 per cent (Table 31), and is actually below the median of the normal distribution of such qualities a distance of −0.14 standard deviation unit. Specimen E was rated as better than A only 11.1 per cent of the time. It therefore is 38.9 per cent below the median quality represented by A. Table X in Garrett, or Table 51 in Thorndike, indicates that a specimen of such quality lies 1.22 sigma units below the quality of Specimen A. All the sigma-unit values given in Table 32 were obtained from these tables in a similar manner.

Table 32 gives the standard deviation distance between each sample and all other specimens not at the median or at the extreme end of the scale. These data now make it possible to compute the scale difference of the samples. The formula [3] for obtaining the scale differences is as follows: $S_{s\sigma} = \frac{\sqrt{2}\cdot x1k - x2k}{N}$, in which $S_{s\sigma}$ equals the scale separation of the samples in sigma units, $x1k - x2k$ equals the sum of the sigma-unit differences for the two specimens, and N is the number of such differences.

Example: Differences of Samples F and H. They rank third and fourth in the series of twenty-two samples of soldering. (See Table 29.)

F	H	Sigma-Unit Differences
−1.22	−1.22	0
+0.76	+0.14	0.62
−1.22	−0.76	0.46
−1.22	−1.22	0
−1.22	+0.76	1.98
−0.76	−0.76	0
−1.22	−0.76	0.46
+1.22	+1.22	0

Scale differences $= \frac{\sqrt{2}\cdot 352}{\circ} = \frac{1.414\cdot 352}{8} = \frac{4.9772}{8} = 0.62.$

[3] This formula and the proof for same are given by Thurstone in *Journal of General Psychology*, Vol. 1 : 405-423, 1928.

TABLE 32

SIGMA DIFFERENCES EXPRESSED AS DEVIATIONS FROM THE MEDIAN (50 PER CENT)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
A	X		− .14		−1.22		− .14		+1.22	− .76				−1.22		+ .76	+ .14	+ .76	− .14		+ .14	
B		X							−1.22							− .76		−1.22			−1.22	
C	+ .14		X		+ .43	−1.22	+ .43	−1.22		+ .14	− .43	+ .14	−1.22	+ .14	− .76	+ .76	+ .43	+1.22	+ .43		+ .76	− .43
D			−1.22	X		+ .76		+ .14			+ .76		+1.22									+1.22
E	+1.22		− .43		X		+ .76	−1.22	+1.22	+ .14		− .43	− .76	− .14	−1.22		+1.22	+ .76	+ .43		+ .76	− .43
F			+1.22	− .76		X		− .14			+1.22	+1.22	+1.22	+ .76	+1.22					−1.22		
G	+ .14		− .43		− .76		X		+1.22	− .43		−1.22	−1.22	−1.22			− .14		− .43		+ .76	−1.22
H			+1.22	− .14	+1.22	+ .14		X		+1.22	+ .76	+1.22	+ .76	+ .76	+ .76		+1.22			−1.22		+ .76
I	−1.22	+1.22			−1.22		−1.22		X					−1.22		− .14	− .43	−1.22	−1.22		−1.22	
J	+ .76		− .14		− .14		+ .43	−1.22		X		− .43		− .14	−1.22	+1.22	+ .43	+ .76	+1.22		+1.22	− .76
K			+ .43	− .76		−1.22		− .76			X	+1.22	+ .14	+ .76	− .14							
L			− .14		+ .43	−1.22	+1.22	−1.22		+ .43	−1.22	X	− .76	− .14	−1.22	+1.22		+1.22				+55.6
M			+1.22	−1.22	+ .76	−1.22	+1.22	+ .76			− .14	+ .76	X	+ .43	− .43	+1.22	+1.22	+1.22	+1.22		+1.22	+ .43
N	+1.22		− .14		+ .14	− .76	+1.22	− .76	+1.22	+ .14	− .76	+ .14	− .43	X	−1.22		+ .76	+1.22	+ .76		+1.22	− .14
O			+ .76		+1.22	−1.22		− .76		+1.22	+ .14	+1.22	+ .43	+1.22	X							+1.22
P	+ .76	+ .76	− .76						+ .14	−1.22		−1.22	−1.22			X	−1.22	− .43	− .76		− .43	
Q	− .14		− .43		−1.22		+ .14	−1.22	+ .43	− .43			−1.22	− .76		+1.22	X	+ .14	+ .14		+ .14	
R	− .76	+1.22	−1.22		− .76				+1.22	− .76		−1.22	−1.22	−1.22		+ .43	− .14	X	− .43		− .14	−1.22
S	+ .14		− .43		− .43		+ .43		+1.22	−1.22			−1.22	− .76		+ .76	− .14	+ .43	X		+ .43	
T						+1.22		+1.22												X		
U	− .14	+1.22	− .76		− .76		− .76		+1.22	−1.22			−1.22	−1.22		+ .43	− .14	+ .14	− .43		X	
V			+ .43	−1.22	+ .43		+1.22	− .76		+ .76		− .14	− .43	+ .14	−1.22			+1.22				X

(From Table 51 in Thorndike's Mental and Social Measurements)

The application of the same method to the other samples results in the scale differences shown in Table 33. This table gives the rank order according to the nine judges and the scale differences between each sample on a comparable basis. It is known that the distance from T to D is 1.09 sigma units, and on the same scale the difference between D and F is 0.16 unit.

TABLE 33

SCALE DIFFERENCES BETWEEN SPECIMENS

Samples	Difference in Units
T-D	1.09
D-F	0.16
F-H	0.62
H-K	0.87
K-O	0.25
O-M	0.57
M-V	0.49
V-L	0.31
L-N	0.33
N-C	0.57
C-E	0.48
E-J	0.41
J-G	0.98
G-S	0.29
S-Q	0.51
Q-U	0.66
U-A	0.64
A-R	0.56
R-P	0.59
P-I	0.88
I-B	0.29

96. Establishing a Point of Origin for the Scale.

The next problem is to establish a zero point. It is permissible in a quality scale of this type to assume a zero point. According to the rating of the nine judges, sample T was rated the poorest of all the samples. The question then arises: Is sample T the poorest conceivable soldered lap joint? The answer is that it is very poor, but could be worse and still hold together. Therefore, for the purposes of this scale, Specimen T is assumed arbitrarily to have a value of 0.9 unit above zero. There probably could not be a soldered lap joint of zero quality since it would not hold together at all and could not be considered a soldered joint. In assuming a zero point for a quality scale in industrial education, it seems advisable to select a point from 0.8

to 1 above zero. After all, such a zero point or point of origin for the scale is arbitrarily set up, whatever the method used.

With the zero point assumed as 0.9 unit, the twenty-two samples are then ranked along the scale by adding the scale differences to the assumed zero according to the relative rankings on the scale. Table 33 shows the sigma-unit differences in quality of each adjoining pair of the twenty-two samples. Table 34 shows the scale values assigned to each specimen. Sample T, being the specimen of poorest quality, is given a scale value of 0.9, on the assumption that it has a quality value of approximately 0.9 unit above the point designated as the arbitrary zero point of the scale. Sample D is 1.09 sigma units better than Sample T; therefore the scale value assigned to Sample D is $1.09 + 0.90 = 1.99$ units above zero quality. Each exercise in ascending order of merit is assigned a scale value corresponding to its merit in relation to the next poorer specimen. The values in the parentheses in Table 34 are sigma units of difference between the pairs of specimens. The ascending values are the scale units of value or merit assigned to each of the twenty-two specimens comprising the scale. A graphic presentation of the relationship of these specimens to the scale and to each other is given in Fig. 18.

After the samples are ranked and scaled, the final step is to select samples suitable for use in the quality scale. In doing this, the au-

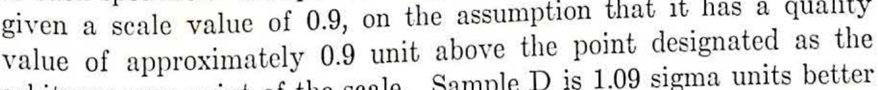


FIG. 18.—Specimens Assigned to a Linear Scale (Zero assumed to be .9 step below sample "T").

thors have found that from eight to twelve samples make a satisfactory scale for checking quality in industrial education. The first object is to select samples for the scale whose quality values represent approximately equal distances along the scale. This gives the samples a definite rating. The scale can then be used as a measuring instrument for rating quality. Fig. 19 shows a scale with eight samples and another with thirteen samples. Both of these scales are taken from

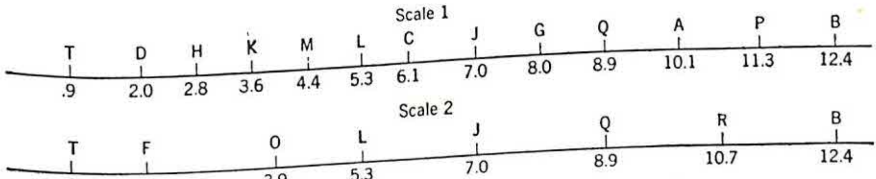


FIG. 19.—Two Quality Scales Selected from the 22 Specimens.

the combined data in Fig. 18 (Table 34). It frequently occurs that several samples will fall very close together on the scale, as *D* and *F* in Fig. 18. This means that two samples of approximately equal quality are available for that point on the scale. This also explains the unnecessary computation involved when a large number of samples

TABLE 34

Scale Values

Zero assumed to be 0.9 step below sample T.

T	0.90	
	(1.09)	
D	1.99	S
F	(0.16)	
	2.15	S
H	(0.62)	
	2.77	S
K	(0.87)	
	3.64	S
O	(0.25)	
	3.89	S
M	(0.57)	
	4.46	S
V	(0.49)	
	4.95	S
L	(0.31)	
	5.26	S
N	(0.33)	
	5.59	S
C	(0.57)	
	6.16	S
E	(0.48)	
	6.64	S
J	(0.41)	
	7.05	S
G	(0.98)	
	8.03	S
S	(0.29)	
	8.32	S
Q	(0.51)	
	8.83	S
U	(0.66)	
	9.49	S
A	(0.64)	
	10.13	S
R	(0.56)	
	10.69	S
P	(0.59)	
	11.28	S
I	(0.88)	
	12.16	S
B	(0.29)	
	12.45	S

of approximately the same difficulty are used, because several samples fall at about the same place on the scale and only one or two are needed for that point to make the quality scale.

Fig. 20 shows photographic reproductions of scales of quality for end-splices, underwriter's knots, solder joints, dados, and flat- and round-head screws.

97. Reliability of Ratings on Quality Scales.

The question may now be asked, are quality scales reliable measuring instruments? Both the reliability of the pooled judgments of experts and the reliability of ratings on quality scales have been shown

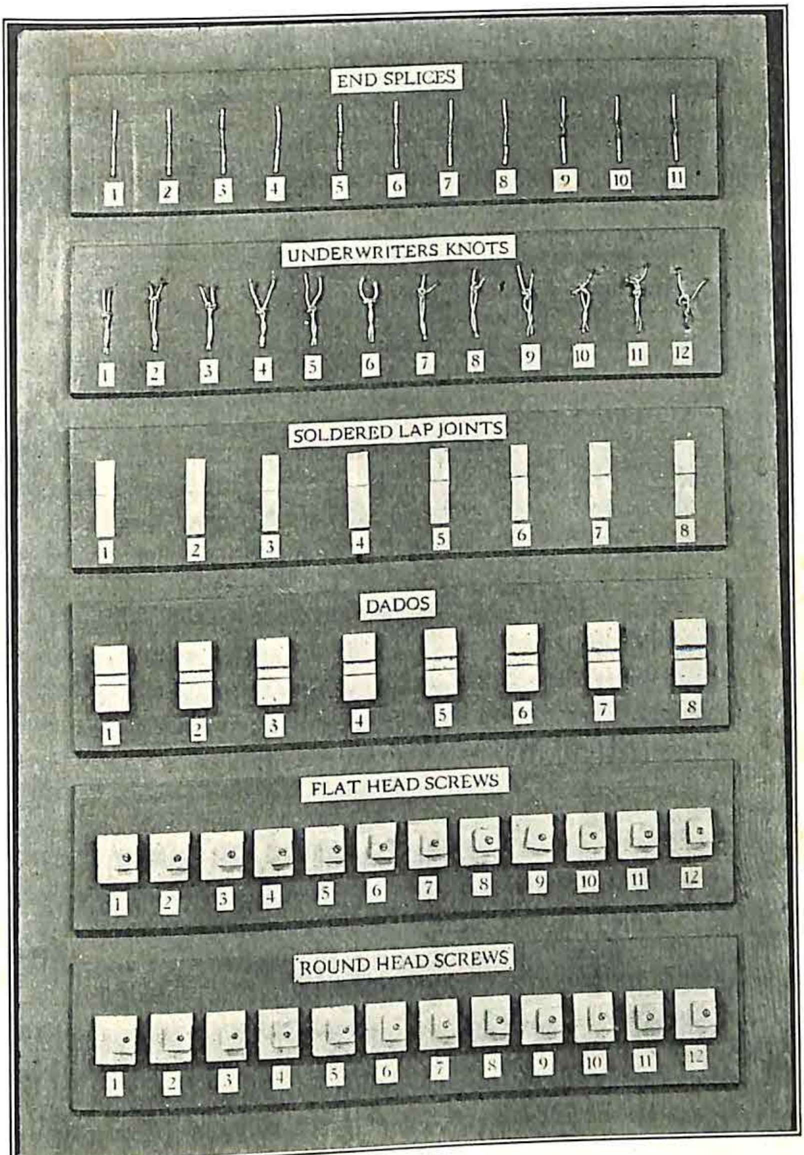


FIG. 20.

to be high by Paterson, Elliott, Anderson, and others [4] in developing a criterion of mechanical ability. These workers found the pooled judgments of ten qualified judges to have a reliability of .90 or better. They also found the reliability of quality scales to be high when the ratings of two or more judges are averaged.

[4] Paterson, Elliott, Anderson, and others, *Minnesota Mechanical Ability Tests,* The University of Minnesota Press, Minneapolis, Minnesota, pp. 194-202.

The authors have found that the reliability of rating of shop products by means of quality scales is much higher than that normally obtained under subjective conditions. However, it is advisable to average three or four independent ratings if a reliability of .90 or better is desired. These ratings may be made by three or four qualified teachers or by the same teacher at widely separated intervals.

SUMMARY

The rating of shop and drawing products involves problems which confront practically all industrial education teachers. The subjective ratings of shop and drawing teachers are as unreliable as teachers' marks in other subject-matter fields. The rating of a shop project or a drawing requires the making of a complex judgment involving many variables. The reliability of industrial education teachers' marks can be improved by using a project rating scale.

Shop and drawing projects are rated by inspection, physical measurement, and quality scales. A project rating scale lists the variable characteristics of the project upon which the judgment is based so that they may be considered individually, and the total rating is the sum of the individual ratings. The project rating scale is also a valuable teaching device for developing appreciation of those factors which produce quality in workmanship and in diagnosing individual pupil difficulties.

The development of a quality scale involves the collection of representative specimens of varying merit, the arrangement of these samples in order of merit by means of the use of pooled judgments, the calculation of the differences in quality of the specimens in terms of sigma units, the establishment of a zero point, and the arrangement and evaluation of the specimens along the scale in relation to one another and to the point of origin of the scale. The technique of pooled judgments is probably sufficiently reliable for most purposes when as many as ten or more qualified and critical judges are used. Quality scales are reasonably reliable measuring instruments, but reach their highest efficiency when the ratings of three or more judges are averaged.

SUMMARY EXERCISES FOR DISCUSSION

1. Point out the distinguishing features of project rating scales and quality rating scales for shop products.
2. Make a project rating scale for some industrial art field other than mechanical drawing. The mechanical drawing rating scale on pages 153 to 155 may be followed as an example.
3. Recapitulate the steps in preparing a quality rating scale.

4. In securing specimens for the construction of a quality rating scale, why is it essential that a wide range of quality be sampled?
5. What importance do you see in the establishment of the zero point of a quality scale?

SELECTED REFERENCES

Ericson, Emanuel E., "Grading Shopwork," *Industrial Education Magazine*, Vol. 28: 227-228, January, 1927.

Hager, Carl J., "A Systematic Method of Grading Shop Projects," *Industrial Arts Magazine*, Vol. 18: 376-378, October, 1929.

Kelly, Truman L., *Statistical Method.* New York: The Macmillan Company, 1923.

Manger, Emerson W., "Grading Drawings," *Industrial Education Magazine*, Vol. 24: 116-117, October, 1922.

Paterson, Elliott, Anderson, *et al.*, *Minnesota Mechanical Ability Tests.* The University of Minnesota Press, 1930, pp. 188-202.

Thurstone, L. E., *Journal of General Psychology*, Vol. 1: 405-423, 1928.

CHAPTER XIII

RATING AND DEVELOPING PERSONALITY AND CHARACTER TRAITS

98. Importance of Character and Personality Traits.

The very commonplaceness of the terms character and personality may account for the fact that they have been given no very exact definition. In a vague way we know what character is, and in a similarly vague way we speak of personality, yet probably no other single outcome of our social or educational programs is so important as the development of proper character and personality traits. Every serious-minded teacher, parent, or pupil realizes the importance of a good character and a pleasing personality. But what is character? What is personality? Can they be developed, or are they inherent? Are we mere slaves of heredity doomed to act and to think in the same way in similar environmental conditions, or is our destiny to a certain extent in our own control?

Character has been defined in a discussion of psychology for teachers as "the sum total of his [*an individual's*] behavior in relation to the world about him and to his fellow beings."[1] Thus, character is taken to be synonymous with the behavior aspects of personality. Character is revealed by an individual's responses to stimulation. If the responses are those commonly considered pleasing or acceptable to his fellows, the individual is described as having a "good" character. If the reverse, he has a "weak" character.

Lay usage has tended to attach to personality the notion of *individuality* or *distinctiveness* in character. In many respects this has had an unfortunate effect, for in their efforts to achieve distinctiveness many individuals have revealed phases of character which are far from pleasing. Individuality is undoubtedly important but not so important that it should be achieved at the expense of honor, honesty,

[1] Benson, C. E.; Lough, J. E.; Skinner, C. E.; West, P. V., *Psychology for Teachers*, Ginn and Company, Boston, 1926.

morality, bravery, modesty, or other attributes of character, which, in the past at least, have been considered desirable.

Much damage has been done by self-styled "psychologists" who have filled the literature of today with high-sounding suggestions by which an individual is to find his latent powers and suddenly develop a strong and forceful personality. Valentine [2] states that a "little exaltation will hurt no one. It is a healthful sort of feeling, but is no substitute for intelligence, vocational ability, moral habits, leadership, culture, social habits, or any other desirable quality which inheritance and creative experience alone can supply." The school, the shop, and the home face a most difficult and critical problem in providing the right kind of stimuli for the development of desirable character and personality traits.

Personality can be modified and improved through diligent effort over a period of time, but it is not a simple task which can be accomplished in a few weeks or months. Personality needs always to be modified and developed in the light of changing social and vocational conditions. Obviously, since our personalities have resulted from the modification of our original natures by our environment, we can consciously develop more desirable modes of conduct by selecting the type of adjustments and then developing habits accordingly. It is true that too often personality emerges with little conscious knowledge on the part of the pupil of what the desirable traits are which have been found essential to success in life. However, the pupil and teacher could at least in part direct the development of personality if they have in mind an ideal toward which to work.

Jones [3] states that "In general, personality is the same as individuality: it is that group of qualities and characteristics that makes one an individual, that set him off from other individuals. As such, it is the sum total of abilities, skills, interests, and physical and mental characteristics that he possesses, or better still the combination of all of these."

As an illustration, let us consider the modern automobile. Its present state of development is the partial realization of an ideal of transportation which has been taking shape during a third of a century. It required much planning and testing to bring the automobile to its present state of efficiency. Today we use adjectives such as the following to describe its traits: durable, economical, safe, speedy, com-

[2] Valentine, P. F., *The Psychology of Personality*, D. Appleton and Company, New York, 1927, p. 355.

[3] Jones, Arthur J., *Principles of Guidance*, McGraw-Hill Book Company, New York, 1930, Chapter X, p. 148.

fortable, and dependable. The characteristics of the automobile are constantly being refined so that they can better meet the changing social and economic conditions of the times. As long as automobiles are used there will be a constant need for refinement and adaptation. The same is true of a human personality. It is the result of the interplay of many variable factors, and even after it is well developed it needs constant refinement in order to keep adjusted to changing environmental conditions.

99. Measuring Personality and Character Traits.

Devices for the measurement of personality and character traits have generally taken the form of rating scales rather than of tests. However, a number of these devices are so arranged that the individual records his own reactions quite objectively. In this respect they resemble tests. Such instruments differ from the typical rating scale also in that the individual responding to the exercises is frequently not aware of the fact that he is being measured for any particular quality or trait. For example, in the *Bernreuter Personality Inventory*[4] the subject is asked to indicate his reaction to 125 questions by encircling one of the answers Yes, No, or ? which precedes each question. The following samples taken from the test itself will illustrate the types of exercises used:

1.	Yes	No	?	Does it make you uncomfortable to be "different" or unconventional?
2.	Yes	No	?	Do you day-dream frequently?
5.	Yes	No	?	Do you ever give money to beggars?
15.	Yes	No	?	Do you usually object when a person steps in front of you in a line of people?
25.	Yes	No	?	Do you study the motives of other people carefully?
50.	Yes	No	?	Do you usually try to avoid arguments?
100.	Yes	No	?	Do you prefer to be alone at times of emotional stress?

By the arrangement of the material in the test several different aspects of personality are measured at one time. According to the author, the scales used in the scoring of the responses to the test are very reliable. This may be due in part to the fact that the traits measured are not readily detectable from the test itself. The significance of the individual's response to the questions is brought out by the use of four separate scales in the scoring of the answers. For example, Scale B1-N is a measure of neurotic tendency. Persons who

[4] Bernreuter, Robert G., *The Personality Inventory*, Stanford University Press, Stanford University, California, 1931.

score high on this test tend to be emotionally unstable. In the case of the exercises used in the sample above, a person who answers question 1 with "Yes" scores +2 points. A "Yes" on question 2 adds five more points. On the other hand a "No" for question 5 deducts 6 points. When Scale B2-S, the scale for self-sufficiency, is used on these same answers, however, a reply of "Yes" for question 1 gives a score of −4; a "Yes" on question 2 gives a score of +1; a "No" on question 5 gives a score of −3 points; etc.

The remaining scales, B3-I for introversion-extroversion, and B4-D for dominance-submission, are applied in a similar manner. Scoring the individual's response to the questions by each of these four scales gives rise to four sets of personality scores for each of which norms are available. Persons scoring high in self-sufficiency are the types who prefer to be alone, do not seek sympathy or encouragement, and tend to follow their own inclinations rather than seek the advice of others. Persons scoring high on the introversion-extroversion scale are inclined to be imaginative. Those scoring low on this scale rarely worry and prefer to act rather than to dream. Persons scoring high on the dominance-submission scale tend to dominate others in face-to-face situations.

Analysis of personality such as is afforded by the *Bernreuter Personality Inventory* has been used with success and considerable reliability with high-school students, college students, and adults. The inventory itself is self-administering, there are no time limits, and each person interprets the questions for himself.

Another type of attempt to secure an unbiased picture of certain personality traits without the subject's being completely aware of the traits on which he is to be measured is represented in the *Loofbourow-Keys Personal Index.*[5] According to the statement in the manual for the test itself, it "is an instrument for the detection of attitudes indicative of problem-behavior. It is intended for use in group surveys to identify those boys whose personal and social maladjustment is such that they are, or are in danger of becoming, serious disciplinary problems." It is standardized for use in the junior-high-school grades, although it has been found useful in senior-high-school and in continuation-school groups. Brief samplings from each of the four test parts comprising the battery are given here for illustrative purposes. The total number of exercises in each test is indicated in the samples.

[5] Loofbourow, Graham C., Keys, Noel, *Personal Index;* Educational Test Bureau, Inc., Minneapolis, 1933.

TEST 1

Directions: This is a test of your word knowledge. Put an X in front of each word you know. There are 100 words.

—perceive
—restore
—grole
—luxury
—rettle
—verify
—proportion
—galine
—exceed
—patient

TEST 2

Directions: Below are some words and phrases with some statements about each one. Mark an X in front of the one statement under each word or phrase which tells best how you feel about the thing named. Mark only one statement under each one.
(Seventeen items.)

---

1. Chums:
   —It is hard to go without them.
   —You cannot always trust them.
   —They sometimes squeal on you.
   —They help you lie out of things.
3. Teachers:
   —They work hard.
   —They know they can punish you.
   —They are not fair to you.
   —They are kind of cranky.
10. Policemen:
   —They have it in for the kids.
   —They are glad to help you out.
   —It is fun to fool them.
   —They are just big bluffs.

TEST 3

Directions: Read carefully and underline the one response which makes the best answer for you. Underline only one.
(Twenty-one items.)

---

1. Do you call another person by a nickname he or she does not like?	Almost always	Sometimes	Hardly ever
2. Do you keep right on studying when the teacher goes out of the room?	Almost always	Sometimes	Hardly ever
21. Do you speak pleasantly to all the people you know even if you do not like them?	Almost always	Sometimes	Hardly ever

Test 4

Directions: Answer every question as truthfully and honestly as you can by drawing a line under the right answer, as shown in the samples.

A. Do you eat more than once a week?	Yes	No
B. Would you rather have a dime than dollar?	Yes	No

(Eighty-nine items.)

---

1. Would you like to wear expensive jewelry, rings, etc.?	Yes	No
2. Do you feel bored a good deal of the time?	Yes	No
50. Are you anxious to get away from school and get a job?	Yes	No
89. Do you know anybody who is trying to do you harm or hurt you?	Yes	No

The test is constructed in such a way that the undesirable responses are the ones scored. The "problem" responses were determined by comparing the answers given by "problem" boys with those made by others of the same age and intelligence. Test 1, False Vocabulary, is scored by allowing 1 point for each fictitious word. The possible score on this part is 30 points. In Test 2, Social Attitudes, three of the four possible answers are socially unacceptable. Each answer so marked counts one error. The score is the number of errors multiplied by 2. The possible score is 34 points. Test 3, Virtues, is scored by allowing 1 point for each fault confessed. The score is obtained by multiplying the number of confessed faults by 2. The possible score is 42 points. For Test 4, the Adjustment Questionnaire, the score is the number of "problem" responses. The possible score is 89 points.

The sum of the four scores listed above gives the subject's personal index. The highest possible index is 195 points. The author's statement of the significance of these personal indices is quoted from the examiner's manual for the test:[6] "An index of 30 or less is clearly insignificant as regards problem behavior . . . A score of 40 or higher, however, strongly suggests an unwholesome trend, since such scores are made by three out of four reform school boys, as compared with only one in five of the others. Scores of 50 to 60 are much more highly indicative, and occur but rarely in unselected groups.

"By noting those boys who show high personal indexes, say 40 or over, principals, teachers, and counselors will have early brought to their attention those individuals most likely to become disciplinary problems and presumably in gravest need of observation and counsel."

The reliability of the battery is approximately .90.

[6] *Op. cit.*

100. Rating Scales for Character and Personality Traits.

In addition to these two types of more or less objective tests designed to reveal personality and character differences a number of general rating scales are also in common use. These are of two general types. Individuals are ranked according to their standing in regard to the specified character traits, or the character traits are rated and assigned a rank. The best rating scales tend to lessen the spread of teacher judgment and when two or three judgments are averaged are found to give fairly reliable estimates for an individual's traits. Hollingworth,[7] Shen,[8] and Rugg[9] have all reported studies on the reliability of rating character traits. On an average the reliability of these ratings is about .55 but varies from .40 to .70 on the best rating scales depending on the traits rated. Whenever possible in using rating scales it is desirable to have two or three teachers rate the same pupil and then average the ratings. Rugg reports this to be a fairly satisfactory method although it is at times difficult to get pupils rated by three different individuals who know the pupil equally well.

Self-rating scales are also used for allowing pupils to rate themselves. This practice has some value in calling the pupil's attention to desirable traits, and in giving a better understanding of some of the desirable and undesirable traits. It has been found that pupils tend to rate themselves too high on the desirable traits, but in general the reliability of their ratings is about the same as results obtained on other types of rating scales.

Industrial education teachers have a definite need for measures of character and personality traits. Since the scales are the best measures available, it seems desirable to use them as one aid in pointing out and rating character and personality traits of pupils. It is well to bear in mind, however, that the general reliability of the best scales is only about .55, and the results obtained are only suggestive but are to be preferred to the unaided subjective judgment of the teacher.

The most common use for rating scales in teaching is to study pupils who are doing unsatisfactory work, although the results are valuable in developing character and personality traits in all pupils.

[7] Hollingworth, H. L., *Judging Human Character,* D. Appleton and Company, New York, 1922.

[8] Shen, E., "The Reliability Coefficient of Personal Ratings," *Journal of Educational Psychology,* Vol. 16: 232-36, April, 1925.

[9] Rugg, H. O., "Is the Rating of Human Character Practicable?" *Journal of Educational Psychology,* Vol. 12: 425-38, 485-501, November, December, 1921; 13: 30-42, 81-93, January, February, 1922.

Scales have also been developed for rating shop teachers. Scales of this type can be used by the supervisory officers or for self-rating and analysis. Teachers and supervisors should keep in mind at all times in using the results of rating scales that the average or median of several ratings is more reliable than the rating by one individual.

The values of trait rating are summarized as follows by Dr. Hughes: [10]

1. Trait rating affords the teacher a better understanding of the individual student. The teacher cannot conscientiously fill out the record unless she knows.

2. It affords a basis for the modification of school and classroom procedures. If these traits and attitudes are valuable in education, then the school situations and methods need to be adjusted to their development.

3. It gives a better understanding of special groups, such as above-average and superior students who are doing poor school work, or below-average students who are doing superior work, which entitles them to membership in honor societies, etc.

4. Follow-up of trait rating brings out the fact that teachers' marks for scholastic achievement are based, to a large extent, on the student's possession of desirable character traits. These data indicate that teachers should be trained to give marks for scholastic achievement alone, and that other marks should be devised for character traits and attitudes, because they are important enough to deserve separate consideration.

5. Cooperation of parents in filling out trait-rating scales for their own children will tend to bring about a better understanding between the home and the school, resulting in better cooperation.

6. Self-rating by the students, on the same scale upon which they are being rated by teachers and parents, will tend to turn the students' attention to the importance of cultivating proper traits and attitudes. Students are inclined to attach importance to things which are being measured, recorded, and used.

7. Justice in marking and teacher judgments will be more apt to be accorded all groups of students when teachers and counselors have a more accurate knowledge of the character traits of their students than they could possibly gain by their own subjective judgments.

8. Trait rating and analysis will result in more scientific counseling, because it will help to furnish a wider basis of knowledge and information about the students upon which to predicate advice.

101. Desirable Traits in Industrial Education.

It is generally agreed that the public school has responsibilities in developing and helping to establish desirable character and personality traits. The industrial education teacher and his co-workers in other instructional fields have a share in this responsibility. We believe, also, that they have a definite contribution to make. In the first place, it requires time to develop personality. Personality continues to de-

[10] Hughes, W. Hardin, "Organized Personnel Research and Its Bearing on High School Problems," *Journal of Educational Research*, Vol. 10: 386-398, December, 1924.

velop as the result of an interplay between the original nature of the individual and the environment, regardless of the conscious attention given to it. Accordingly, the problem of the industrial education teacher who wishes to help develop desirable personalities in his pupils is first to find out what types of personalities adjust themselves best in our present complex social and economic life. In the second place, a desirable personality must be cultivated, and the desirable traits must be encouraged. The undesirable ones must be weeded out and their expression discouraged. However, it must also be remembered that personality development is limited by the innate capacity of the individual, and so its development may be expected to vary markedly under similar environmental conditions.

102. Constructing and Using Scales for Rating Personality Traits in Industrial Education.

Although scales for rating personality traits have not been as widely used as tests of information, intelligence, and mechanical aptitude, several usable techniques have been developed which will aid the teacher who desires to construct and use the trait rating scales. The first problem in the construction of such scales is the selection of the traits to be rated. This selection may be based on the observation and experience of the teacher, conferences with other instructors, conferences with the administrative officers, talking with the pupils and their parents, conferences with industrial and social leaders in the community, and suggestions from authoritative literature. After a rather exhaustive list has been prepared, a number, perhaps twenty, of the most important items should be selected for use in the rating scale. This selection may be accomplished through the use of pooled judgments of teachers and others interested in personality development. The traits selected must also conform to the general purpose of the course of study and be of such a nature that the industrial education teacher will have an opportunity to observe the expression of the traits in and about the school. For purposes of illustration let us consider the following personality traits which were selected by the authors after an analysis of the problem. The traits are listed and defined in terms of observable pupil responses.

Self-reliance. This means that there has been developed in the student the habit of planning tasks carefully and thoughtfully and of carrying them out with only necessary assistance. The problem is obviously too difficult before assistance is called for by the student.

Industry. This means a habit of careful, thoughtful work without loitering or wasting time.

Readiness to assume responsibility. This means that a task though difficult should not be avoided if worth doing, and when once undertaken should be carried through to completion.

Punctuality. This means the ability to arrive on time and fit oneself to a program.

Cooperation. This means a readiness to assist others when they need help, and to join in group undertakings.

Consideration of others. This means a thoughtful attitude in the making of things easy and pleasant for others. It involves keeping things in order, putting tools away in good condition, and always doing a full share of work where others are involved.

Cleanliness and neatness. This means the ability to keep physically clean and neat in both work and dress.

An optimistic viewpoint toward life. This means an appreciation of the joy of living and a belief that life is worth while.

After the character traits to be rated have been selected and defined, the next step is to put them into a rating scale which will permit the greatest amount of objectivity in scoring. There are several acceptable methods of accomplishing this, depending on whether the pupils are to be given a relative rank according to their traits, or whether the character traits of individuals are to be rated and assigned a rank. Dr. Hughes [11] gives the following three procedures used in rating:

Method I. Normal Distribution. In this method we apply the principle represented in the "normal curve of distribution." In any large number of unselected cases we find a few who possess a given quality in maximum degree, and a correspondingly small number who possess it in minimum degree. A much larger number, however, possess the quality in average degree. This general principle holds whether we consider height, weight, strength, or any other measurable quality or characteristic. For a scale consisting of five equal steps, we should have approximately the following distribution of cases on a percentage basis:

Lowest	Inferior	Medium	Superior	Highest
7	24	38	24	7

But for practical purposes we have adopted a theoretical distribution as follows:

Lowest	Inferior	Medium	Superior	Highest
10	20	40	20	10

Assuming that the individuals who are to be rated are unselected and representative we should have 10 in 100 marked "highest"; 10, "lowest"; 40, "medium"; and 20, "inferior" and "superior" respectively. A convenient method of rating such a group is to have the names on individual cards and then arrange these cards in five piles according to the percentage distributions required.

[11] Hughes, W. Hardin, "General Principles of Rating Trait Characteristics," *Educational Research Bulletin,* Pasadena, Vol. 3, Nos. 5 and 6, February-March, 1925.

The rater should as far as possible dismiss from mind every other item of the scale and concentrate on the one being rated.

The method of "normal distribution" is most usable with large and unselected numbers. When the number of cases is small and selected the method is defective. For this reason another method, based on the same principle, is presented.

Method II. The Master Scale. To use this method, proceed somewhat as follows: Suppose the trait for which a master scale is to be made is industry.

1. Recall any student known to possess this trait in highest degree. Write his name opposite "highest" in the master scale.

2. Now, recall any student known to possess this trait in lowest degree and write his name opposite "lowest" in the master scale.

3. Then recall any student known to possess the trait in average degree, write his name opposite "medium."

This gives three definite standards for comparison. The other places in the scale may now be filled in with names of two students half way between "medium" and "highest" and half way between "medium" and "lowest," respectively. You now have a master scale as follows:

Master Scale for Industry

Rating	Person	Numerical Value *
Highest	John Jones	180
Superior	Dick Brown	140
Medium	Sam Johnson	100
Inferior	Henry James	60
Lowest	Bill Smith	20

* The numerical values here assigned represent the half-way point in each 40 of a 200-point scale.

With this master scale in hand, the teacher is now ready to rate her students in industry. Suppose Tom Black is to be rated. The teacher quickly decides whether Tom is as good as John Jones, as poor as Bill Smith, or just about like Sam Johnson, etc. Master scales for the other traits may be made and used in the same way.

The advantages of this scale are that it is objective and that small numbers of students can be rated without immediate reference to the "normal curve of distribution." In the long run, however, the percentage distributions should approximate those given under Method I.

103. A Useful Personality Rating Scale.

The authors have found the following type of scale valuable in rating personality traits in industrial education classes. It will be noted that it uses a form of the graphic method with the quality units spaced roughly corresponding to the normal distribution curve.

A Graphic Rating Scale

for Personality Traits in Industrial Education

Name______________________________ Date____________

Instructor____________________________ Rating____________

Directions: Provision is made for two or more ratings of each personality trait on the basis of observable pupil responses. Place a check on the line which in your judgment represents a true estimate of the present status of the trait being rated.

Minimum	Average	Maximum

Self-Reliance

Does the pupil plan his work carefully and thoughtfully?

Does the pupil conduct the work with only necessary help?

Does the pupil ask for help when the problem is too difficult?

Industry

Is the pupil in the habit of doing careful and thoughtful work?

Does he loiter or waste time in his work?

Readiness to Assume Responsibility

Is the pupil willing to undertake a worth-while task even though it is difficult?

Does the pupil finish all his work?

Punctuality

Does the pupil arrive on time to classes?

Does the pupil hand his work in on time?

Cooperation

Does the pupil help others when help is needed?

Is the pupil active in group undertakings?

Consideration of Others

Does the pupil have the habit of making things pleasant for his classmates?

Does the pupil help keep the shop in order?

Does the pupil put tools away in the right places?

When the whole class is involved in some work does he do his share or skip away?

Cleanliness and Neatness

Does the pupil wash clean?

Does the pupil dress neatly and in good taste?

Is the pupil neat in doing his work?

Optimistic View of Life

Does the pupil have a natural likable smile?

Does the pupil complain about his lot in life?

Is the pupil liked by his classmates?

To secure a total rating, score each question on the basis of the following key:

1	3	7	13	19	23	25	26

Methods of rating personality traits and of ranking pupils have been given and illustrated, but thus far self-analysis by the individual has not been discussed except to mention that it is not significantly higher in reliability than ratings of traits by outsiders. There is a tendency for individuals to rate themselves too high on the desirable traits and too low on the undesirable traits, just as there is a tendency of persons rating others they know very well to be influenced by the "halo effect." This refers to the tendency of the one rating character traits to rate them all about the same, depending on the rater's general opinion of the individual.

SUMMARY

This chapter presents a number of suggestions for the rating of personality and character traits. Personality in its broader sense is developed in an individual by the interplay of stimuli from the environment and the native capacities of the individual. The thesis of this chapter is that a desirable personality can be developed over a period of time through careful practice and an understanding of what is desirable in a well-rounded personality. Industry, cooperation, consideration for others, self-reliance, readiness to assume responsibility, and an optimistic view toward life are suggestive of desirable traits to be developed by industrial education teachers as co-workers with teachers in other instructional fields.

Tests of personality traits have not proved as reliable as rating scales, and for this reason major emphasis is placed on methods of developing and using rating scales. Authorities report the average reliability of rating scales at around .55. The most reliable ratings are obtained when the ratings of three or more judges are pooled. Some rating scales rank individuals from highest to lowest according to their personality traits; others give a graphic rating of the individual traits. In still other types the pupils are allowed to rate themselves. None of these methods is highly reliable, but they are better than the unaided subjective judgments of teachers.

SUMMARY EXERCISES FOR DISCUSSION

1. Distinguish clearly between the terms character and personality.
2. What, in your opinion, is the responsibility of the industrial arts teacher for the development of desirable personality and character traits?
3. Secure from your instructor a copy of the *Bernreuter Personality Inventory,* administer the exercises to yourself, and prepare a self-analysis on the basis of the personality qualities identified in this instrument.
4. Prepare a personality rating scale including the traits which in your judgment (coupled with information gained from reading in the field) are essential to success in teaching.

SELECTED REFERENCES

Allport, F. H. and G. W., "Personality Traits: Their Classification and Measurements," *Journal of Abnormal and Social Psychology,* Vol. 16: 6-40, 1921.

Book, W. F., *Learning How to Study and Work Effectively.* Boston: Ginn and Company, 1926.

Hartson, L. D., "An Experiment with Rating Scales Based upon a Tentative functional Analysis of the Subjects," Society of College Teachers of Education, *Educational Monographs,* No. XIV, 1925, Studies in Education, University of Chicago Press.

Hughes, W. Hardin, "General Principles of Rating Trait Characteristics," *Educational Research Bulletin*, Pasadena, Vols. 3 and 4, February and March, 1925.

Jones, Arthur J., *Principles of Guidance.* New York: McGraw-Hill Book Company, 1930.

Laird, Donald A., *Increasing Human Efficiency.* New York: Harper and Brothers, 1925.

McKay, H. D. *The Development of Personality Traits.* Ph.D. thesis, University of Chicago, 1930.

Newkirk, Louis V., "The Shop Teacher and Personality," *Industrial Education Magazine,* Vol. 18:370-372, No. 10, October, 1929.

Odell, C. W., *Educational Measurements in High School.* New York: The Century Company, 1930.

Valentine, P. F., *The Psychology of Personality.* New York: D. Appleton and Company, 1927.

Rating Scales and Tests

Cornell, E. L.; Coxe, W. W.; and Orleans, J. S., *Rating Scale for School Habits.* Yonkers, New York; World Book Company, 1927.

Downey, June, *The Will-Temperament and Its Testing.* Yonkers, New York: World Book Company, 1923.

Dragoo, Alva W., *A Rating Scale for Shop Teachers.* Bloomington, Illinois: Public School Publishing Company.

Hughes, W. Hardin, "A Rating Scale for Individual Capacities, Attitudes, and Interests," *The Journal of Educational Method,* Vol. 3:56-65, October, 1923.

McClusky, F. D., and Dolch, E. W., *Study Outline Test.* Bloomington, Illinois: Public School Publishing Company, 1926.

Morris, Elizabeth H., *Trait Index L.* Bloomington, Illinois: Public School Publishing Company, 1929.

CHAPTER XIV

SUMMARIZING THE RESULTS OF TESTING

Experience in the observation of the work of individual students and in the use of tests in the classroom leads to the conclusion that wide differences in pupil accomplishment may be expected. This means that scores representing objective measures of achievement in the classroom will vary widely. Since the human mind is not able to grasp and hold numerous unlike facts in isolation, it is apparent that some fairly simple and accurate methods of summarizing and describing such widely varying results are necessary. This process of summarizing, analyzing, and compressing data so that they may be given adequate description is the application of statistical methods.

Six statistical techniques are very useful in the analysis and interpretation of educational test results. They are: (1) the classification and tabulation of data; (2) the computation and interpretation of the common measures of central tendency; (3) the computation and interpretation of the more common measures of variability; (4) the expression of the extent and nature of the interrelations of measurable factors; (5) the derivation and use of standards, norms, and various derived scores for the purposes of comparison and interpretation of test results; and (6) the use of simple and effective graphic procedures for the presentation of facts. This chapter summarizes the discussion of these six points.

I. THE TABULATION OF TEST SCORES

104. Need for Grouping of Data.

The physical, mental, moral, and social unlikenesses of people make it impossible to describe them in few words. If all men were of the same height it would be a simple matter to describe the height of men. This need for a method of grouping data for convenience in treating and describing the situation is illustrated in the test scores given in Table 35 on page 188. This table shows the scores made by a group of 27 eighth-grade pupils on the *Newkirk-Stoddard Home Mechanics Test*, Form A. An examination of this table shows that Pupils 7 and 9 made the low and the high scores on this test. The re-

maining 25 pupils made scores between these extremes. Even with this small class, with scores ranging from 1 point to 15 points, it is apparent that it is difficult for the teacher to secure a clear picture of the achievement of the class without some treatment of the scores. One of the very easiest of these procedures is the arrangement of the test scores in order of size from the highest to the lowest. This is called *ranking* the scores. These 27 scores have been ranked in descending order of size in Table 36. It enables the teacher to discover readily the highest and lowest scores, and, after some training and experience, to pick the scores which are more or less typical for the group.

The arrangement of test scores in order of size is helpful only when the number of pupils in the class is relatively small, as 10 or less. When more pupils are present it usually becomes necessary to group the scores into the form of a table. This is called *making a frequency distribution,* and the table naturally is called a *frequency table.*

105. Steps in Preparing a Frequency Table.

The three essential steps in the grouping of data in a frequency table are as follows:

1. *The determination of the range of the scores.*

The *range* is the difference between the largest score and the smallest score in the array of series of scores. For the test scores given in Table 35, the range is 13—the difference between 15, the highest score, and 2, the lowest score.

2. *The determination of the number and size of the groups in the classification.*

TABLE 35

SCORES IN NUMBER OF JOBS RIGHT MADE IN EIGHTH-GRADE CLASS ON NEWKIRK-STODDARD TEST OF HOME MECHANICS

Pupil Number	Test Score	Pupil Number	Test Score	Pupil Number	Test Score
1	6	10	10	19	8
2	3	11	7	20	7
3	8	12	6	21	5
4	12	13	6	22	10
5	3	14	9	23	9
6	4	15	5	24	5
7	2	16	7	25	8
8	6	17	11	26	6
9	15	18	6	27	7

TABLE 36

SCORES IN TABLE 35 ARRANGED IN DESCENDING ORDER OF SIZE

Test Score	Pupil
15	9
12	4
11	17
10	10, 22
9	14, 23
8	3, 19, 25
7	11, 16, 20, 27
6	1, 8, 12, 13, 18, 26
5	15, 21, 24
4	6
3	2, 5
2	7

The number of groups, or *class intervals* as they are called, in a frequency table depends upon the range of the scores, as well as upon the size of the interval which it seems best to use. No specific rule covering this situation can be stated. In general it seems safe to say that it is usually unwise to group data into fewer than 15 to 18 intervals. On the other hand, the use of 25 or more intervals may introduce an unnecessary amount of labor in making the tabulation, and many less than 15 may introduce a serious *error of grouping*. Some useful suggestions on the relation of the range to the size and number of the class intervals to use in making a frequency table are given in Table 37.

TABLE 37

SUGGESTED RELATION OF RANGE OF SCORES AND SIZE OF CLASS INTERVALS

For a Range of	Use a Class Interval of
25 or less	1
26 to 69	3
70 to 125	5
126 to 175	7
176 or more	15

Error of Grouping. The so-called error of grouping results from the practice of putting into the same group, or interval, scores which are very widely unlike and in which only a few cases are found. It

is greater when the number of scores is small and when they are found at irregular intervals along the scale. For example, the tabulation of two scores of 44 into a class interval of three units (44, 45, and 46) with a mid-point of 45 involves an average error of grouping of 1.0 point, whereas the tabulation of seven scores of 44, 44, 45, 45, 46, 46, 46 into the same class interval introduces a smaller error. In placing these seven scores in the same class interval it is assumed that they will together have an average value approximately the same as the mid-point of the step (in this case 45). In this latter example, the error is not serious, since the average of these seven scores is actually 45.14 instead of 45. In tables with fewer and larger class intervals, it must be obvious that this error due to grouping may be much greater. The underlying idea in grouping scores into 15 to 25 intervals is to organize the scores into a sufficiently small number of classes so that they may be thought about effectively, and yet not place them in so few groups that important differences are covered up or significant errors of grouping are introduced.

3. *Tabulating the scores.*

The tabulation of test scores corresponds in many respects to the filing of letters or cards. The value or size of the score determines the filing compartment into which it is placed. The necessity for this exact classification makes clear to the student why it is so important that the limits of the class intervals be so exactly established in the preparation of the steps for a table. There must be no question as to the intervals in which each specific score is placed.

Undoubtedly the best way to clear up the problems of tabulation is to illustrate by making use of some actual test scores. In Table 35 the scores made by 27 eighth-grade pupils on the *Newkirk-Stoddard Test of Home Mechanics* are given. These test scores cover a range from 1 to 15 points. The real significance of 27 scores with such a range can not be gathered from the form in which they are given in Table 35. A simple form of grouping or compressing these data is used in Table 36 which shows the number of the individual pupil who made each specific score. As a matter of fact this table is turned into a simple frequency table by the process of setting up a scale of class intervals in place of the actual test scores and by substituting check marks or tabulation marks for the individual pupil numbers. Table 37 suggests that a class interval with a step of 1 unit be used in arrays in which the range is 25 points or less. Accordingly, in this problem, class intervals of 1 with whole numbered mid-points are set up. Table 38 shows the results of setting up the intervals on this basis. The tabulation of the scores is shown in this table in the third

column. In tabulating test scores the common practice is to make a vertical mark (/) for each score of a given magnitude. When the frequency of any given score reaches 5, the fifth frequency is indicated by a diagonal mark crossing four of the vertical tabulation marks. In this manner, the frequencies are conveniently grouped by 5's, which simplifies the summation of frequencies in large populations.

A further illustration of the tabulation of a series of test scores is given in Table 39. The scores used in this case represent total

TABLE 38

DATA FROM TABLES 35 AND 36 ARRANGED IN FREQUENCY DISTRIBUTION

Class Intervals	Mid-Points	Tabulation Marks	Frequencies (f)
14.5–15.5	15	/	1
13.5–14.5	14		0
12.5–13.5	13		0
11.5–12.5	12	/	1
10.5–11.5	11	/	1
9.5–10.5	10	//	2
8.5– 9.5	9	//	2
7.5– 8.5	8	///	3
6.5– 7.5	7	////	4
5.5– 6.5	6	~~////~~ /	6
4.5– 5.5	5	///	3
3.5– 4.5	4	/	1
2.5– 3.5	3	//	2
1.5– 2.5	2	/	1
			Total, or $N = 27$

comprehension scores of a ninth-grade class on the *Iowa Silent Reading Test, Advanced Examination.*[1] The test scores of the 71 students comprising this class are as follows: 104, 129, 94, 87, 118, 146, 109, 163, 140, 125, 58, 86, 102, 103, 133, 77, 117, 114, 99, 110, 103, 93, 123, 137, 89, 118, 117, 107, 109, 117, 114, 162, 135, 115, 101, 109, 150, 130, 100, 109, 140, 102, 110, 148, 94, 122, 139, 115, 105, 125, 104, 141, 127, 100, 107, 116, 136, 142, 96, 103, 111, 145, 99, 105, 108, 98, 126, 112, 152, 114, 109.

The range of the scores is found by subtracting 58, the smallest

[1] *Iowa Silent Reading Test,* Advanced, Forms A and B, World Book Company, Yonkers, New York.

score, from 163, the largest score. The difference is 105. Table 37 suggests class intervals of 5 units for such a range. Accordingly, a class interval large enough to accommodate a score of 163 points is set up at the top of the table. Using a mid-point divisible by the size of the step means that the mid-point of this interval will be 165 and that the limits of the interval will be 162.5 to 167.7 points The

TABLE 39

Comprehension Scores on Iowa Silent Reading Test, Advanced Examination

Class Intervals	Mid-Points	Tabulation Marks	Frequencies (f)
162.5–167.5	165	/	1
157.5–162.5	160	/	1
152.5–157.5	155		0
147.5–152.5	150	///	3
142.5–147.5	145	//	2
137.5–142.5	140	~~////~~	5
132.5–137.5	135	////	4
127.5–132.5	130	//	2
122.5–127.5	125	~~////~~	5
117.5–122.5	120	///	3
112.5–117.5	115	~~////~~ ////	9
107.5–112.5	110	~~////~~ ~~////~~ /	11
102.5–107.5	105	~~////~~ ///	8
97.5–102.5	100	~~////~~ ///	8
92.5– 97.5	95	////	4
87.5– 92.5	90	/	1
82.5– 87.5	85	//	2
77.5– 82.5	80		0
72.5– 77.5	75	/	1
67.5– 72.5	70		0
62.5– 67.5	65		0
57.5– 62.5	60	/	1
			$N = 71$

remainder of the table is developed in a similar way until the complete series of 21 intervals necessary for this range of scores is built up. The table must provide for the entire range of the scores, from the largest to the smallest. The limits of the class interval at the bottom of the table are 57.5 to 62.5, which provides for the score of 58 at the lower end of the range of scores. Table 39 shows the entire

range of the class intervals, the mid-points, the tabulation marks, and the frequencies, based on these 71 reading-test scores.[2]

II. MEASURES OF CENTRAL TENDENCY

The second of the important statistical techniques required in connection with the summary of educational test results deals with the computation and interpretation of the common measures of central tendency. This is the process of computing a single measure or term which may be used in describing the complete array of data in the table. The term *central tendency* arises through the fact that these measures are commonly found near the center of the distributions of scores when the scores are arranged in order of size.

Three measures of central tendency are commonly used in the interpretation of educational tests. These are: *the arithmetic mean, the median, and the mode.* In general, these measures are named in the order of their use in present-day test interpretation. As a matter of fact, the mode is considered to be such an unreliable measure that it is rarely used in educational measurements. In this discussion, consideration will be given only to the first two of the measures of central tendency, namely, the arithmetic mean and the median.

106. The Arithmetic Mean.

Almost everyone knows how to find a simple arithmetic mean or average, as it is commonly called, by dividing the sum of a series of measures by the number of measures. However, not everyone knows that there is a rapid and reasonably accurate method of computing the arithmetic mean of large numbers of measures in frequency tables. The speed and the satisfactory accuracy with which this important measure of central tendency may be computed for distributions of large numbers of cases has made it one of the most popular and useful of the measures of central tendency.

The calculation of the arithmetic mean from a frequency distribution requires a somewhat different concept of the measure than that used when it is computed from ungrouped data. For the 27 test scores given in Table 35 the sum of the measures is 191. The arithmetic mean, the result of dividing 191 by 27, is 7.07. When computed in this way the arithmetic mean does not especially require definition. It is easier simply to state how it is found. When computed by the

[2] A much more detailed discussion of the problem of tabulating test scores will be found in Greene's *Work-Book in Educational Measurements* (Longmans). Extensive practice in tabulation of test scores is given in Problems 1, 2, 3, and 4 of Work-Unit I of the above-mentioned *Work-Book.*

so-called shorter method, *the arithmetic mean is defined as a point on the scale such that the sum of the deviations of the values larger exactly equals the sum of the deviations of the values smaller than it is.* For those who think most clearly in concrete terms this arithmetic mean may be considered as the point in a beam of irregular but increasing thickness at which the fulcrum must be placed to bring it into perfect balance.

The actual computation of the arithmetic mean from a frequency table proceeds on the principle of the mathematical determination of the proper position of the fulcrum of such a beam from data resulting from a trial balance. That is, the beam is suspended on the fulcrum as nearly as can be determined by estimation; then the actual amount that the beam is out of balance is measured. Experience shows that the fulcrum must be moved in the direction of the heavy end of the beam in order to bring it into balance. The exact amount of this shift in position depends upon the difference in the forces bearing on the two ends of the beam. If there are 60 units of unbalanced force tending to turn a beam in a certain direction and there are 40 measures (scores) contributing to the distribution, it means that the fulcrum must be moved toward the heavy end of the beam an amount equal to 1.5 scale units of length ($60 \div 40 = 1.5$). This should bring the beam into balance.

107. Steps in Computation of Mean.

The specific steps to be taken in the computation of the arithmetic mean by the shorter method are as follows:

Step 1. Select the mid-point of some central step on the scale as an *assumed mean.* Call this point *zero.* (In computing the arithmetic mean from a frequency distribution the scores in a given step are all assumed to be grouped at the exact center of the step, hence the assumption of the mid-point of the step as the *zero*).

Step 2. Mark off steps of deviation above and below this assumed zero point, maintaining the algebraic signs.

Step 3. Multiply the frequency in each step by the deviation of the step. Carry the algebraic signs of these deviations. Those above the zero step should be plus; those below it should be minus.

Step 4. Find the algebraic sum of these deviations, keeping the sign of the result.

Step 5. Divide this value by the number of cases in the distribution, and multiply this result by the number of units in each step. This result is the *correction* (c).

Step 6. Depending upon the sign of this correction (c), increase or

decrease the value of the mid-point of the step taken as the zero by the amount of c. This should give the *arithmetic mean.*

This procedure may be made clear by actually working the mean of the test scores tabulated in Table 38. The work is shown in detail in the accompanying table (Table 40).

TABLE 40

DISTRIBUTION OF SCORES TAKEN FROM TABLE 35 AND THE CALCULATION OF THE ARITHMETIC MEAN OF THE 27 SCORES

Class Intervals	f	d	fd
14.5 (15) 15.5	1	+7	7
13.5 (14) 14.5	0	+6	0
12.5 (13) 13.5	0	+5	0
11.5 (12) 12.5	1	+4	4
10.5 (11) 11.5	1	+3	3
9.5 (10) 10.5	2	+2	4
8.5 (9) 9.5	2	+1	2
			(+20)
7.5 (8) 8.5	3	0	
6.5 (7) 7.5	4	−1	−4
5.5 (6) 6.5	6	−2	−12
4.5 (5) 5.5	3	−3	−9
3.5 (4) 4.5	1	−4	−4
2.5 (3) 3.5	2	−5	−10
1.5 (2) 2.5	1	−6	−6
	$N=27$		(−45)

Step 1. Assumed mean = 8.0.
Step 2. Lay off deviations.
Step 3. Add plus and minus fd's.
Step 4. Find algebraic sum of fd's. $-45+20=-25$.
Step 5. Divide this algebraic sum by N. $\frac{-25}{27}=-0.926$.
Multiply by size of the step. $-0.926 \times 1 = -0.926$.
Step 6. 8.000
0.926
7.074 = Arithmetic mean.

108. The Median.

The simplification of the work involved in computing the arithmetic mean has done much to stimulate its general use in the interpretation of educational test results in place of the median. However, the ease with which the median may be obtained, and the fact that it does not give undue weight to extreme scores as does the arithmetic mean, have made it a popular measure for use in educational measurements.

Some confusion has been created in the minds of students and teachers through a lack of consistency in the methods of computing the median. For many years it was common practice to instruct users of tests to arrange the test papers for a class with the scores in descending order of size, and take the score on the middle paper as the score best representing the achievement of the class. For a long time this score was called the *median*. As a matter of fact, the measure of central tendency obtained in this way from ungrouped data is a crude median, but in order to distinguish it from the true median computed by data in a frequency distribution there is a tendency to call it the *mid-measure*. The *mid-measure* is a counting median found from ungrouped data. The *median* is always computed from grouped data.

Computing the Mid-Measure. By definition *the mid-measure is the score of the middle paper when the number of test papers is odd, and the average of the two scores nearest the middle when the number is even, assuming that the test papers are arranged in definite order of magnitude.* The method of computing this very simple measure may be illustrated by referring to the data given in Table 35. The 27 test scores given in this table arranged in descending order of size are as follows: 15, 12, 11, 10, 10, 9, 9, 8, 8, 8, 7, 7, 7, 7, 6, 6, 6, 6, 6, 6, 5, 5, 5, 4, 3, 3, 2. The mid-measure is found by counting off the scores until the middle paper is reached. In this case, it will be the score on the fourteenth paper, or 7 points. If there were only 26 papers and the high score of 15 were missing, the average of the thirteenth paper from either end of the scale would be used as the mid-measure.[3] Under these conditions the mid-measure would thus be the average of 6 and 7, or 6.5 points.

Computing the Median. The median is defined as *a point on the scale such that exactly 50 per cent of the cases in the distribution are above it and 50 per cent of the cases are below it.* The median is distinguished from the mid-measure by the fact that the former is a *point* on a scale whereas the latter is an actual *score* on a test paper (or the average of the two scores lying nearest the middle paper). The fact that the score on the middle paper of a series is not the same thing as the middle point in the scale of a frequency table of the same scores makes it important that the two types of measures be defined and distinguished in use. It will be a movement in the direction of uniformity of interpretation of test results if the median is always understood as being computed from data in a frequency distribution.

[3] See Problem 11 in Greene, *Work-Book in Educational Measurements* (Longmans), for additional drill on computing the mid-measure.

In the earlier discussion of methods of tabulation, and particularly in the explanation of the computation of the arithmetic mean, it was pointed out that all the measures falling in a given step are assumed to have the value of the *mid-point* of that step. This is necessary since the computation of the arithmetic mean involves the correction of an assumed mean and this correction may take place in either a positive or negative direction. Now, in computing the median a very different assumption is made, and since this point frequently causes considerable difficulty and confusion, the reasons for making it are explained here in some detail. Since the median is a counting measure and is obtained by counting into the distribution until a point is reached which throws one-half of the frequencies below it, it is necessary to assume that all scores assigned to a given step are distributed *uniformly throughout the step*. When working with the median, or measures of a similar character such as percentiles, all scores are assumed to be scattered through the step in this uniform manner. It may help to think of the steps or class intervals as air-tight compartments, and the scores or frequencies assigned to the steps as a volatile gaseous substance which expands and completely fills the compartment, regardless of how many or how few the frequencies may be. If four scores are assigned to a given step, each one of the cases represents one-fourth of the total area of the step. If there are 20 cases per step, each case is considered to represent one-twentieth of the area of the step in computing the median.

To find the median of a series of scores arranged in a frequency table take the following steps:

Step 1. Divide the total number of cases in the distribution (N) by 2 to determine 50 per cent of the cases. (See Table 38, $N/2$ is 13.5.)

Step 2. Beginning at the bottom of the column of frequencies, count up the frequencies as far as possible without exceeding the half-sum ($N/2$). (In Table 38 the frequencies $1 + 2 + 1 + 3 + 6$ equal 13, which is still less than the half-sum, 13.5.)

Step 3. Take the difference between the half-sum and the subtotal. (In this example the difference between 13.5 and 13 is 0.5 point.) This difference shows the number of cases which must be taken from the step in which the median is located. Since there are four cases in the next step, the step with the limits 6.5 to 7.5, one-half a case out of these four cases must be taken. Thus the median is located 0.5/4 or one-eighth of the way through the step. This shows the proportion of the step which must be passed in counting the fre-

quencies in order to reach a point on the test scale such that exactly one-half of the cases lie below it and one-half lie above it.

Step 4. Since the four cases in this step are assumed to be distributed uniformly throughout the step, the fraction 0.5/4, or one-eighth, represents the fraction of the step which must be passed in order to reach the point known as the median. The fraction one-eighth is equivalent to the decimal 0.125. Since this value 0.125 is in steps and the steps in this table are one-point intervals this value must be multiplied by 1. This means that the median is 0.125 unit beyond (above) the lower limit of the step into which this fractional unit is taken.

Step 5. The beginning (lower limit, since the frequencies were counted from the lower end of the distribution) of the step having the four frequencies is 6.5. Therefore the 0.125 unit must be added to the value 6.5. The median thus becomes 6.625. For practical purposes the decimal may be rounded to 6.63.

Step 6. In statistical work of all kinds accuracy is extremely important. It is therefore very desirable to check all computations. The calculation of the median may be conveniently checked by the simple process of adding the frequencies down from the top of the distribution. In this case the interpolation would be 0.875 and would be subtracted from the *top* of the step, or from 7.5.[4]

The actual work of computing the median of the silent-reading scores given in Table 39 is given in detail in Table 41.

109. Uses of the Arithmetic Mean and the Median.

The question which of these very useful measures of central tendency to use in test interpretation frequently arises. In many respects there is not a great deal of choice. Prior to the general adoption of the shorter methods of computing the arithmetic mean the median was very popular. It is simple to compute, and furthermore, is considered especially suitable for test interpretation because of the fact that widely deviating scores do not unduly influence it. On the other hand, the arithmetic mean is now very easily calculated, and for most experimental purposes it appears to be quite important to have each individual score given weight in the results in direct proportion to its magnitude. The greatly increased use of educational tests for ex-

[4] This and a number of other points in the computation of the median are discussed and explained in connection with Illustrations 7, 8, and 9 (pages 35 to 38 inclusive) in the *Work-Book in Educational Measurements* by H. A. Greene (Longmans). Problems 12, 13, 14, and 15 in this *Work-Book* also provide extensive drill on the finding of medians from all types of distributions.

TABLE 41

DISTRIBUTION OF TEST SCORES OF 71 NINTH-GRADE PUPILS. TOTAL COMPREHENSION SCORES FROM IOWA SILENT READING TEST; ADVANCED

Class Intervals	f
162.5–167.5	1
157.5–162.5	1
152.5–157.5	0
147.5–152.5	3
142.5–147.5	2
137.5–142.5	5
132.5–137.5	4
127.5–132.5	2
122.5–127.5	5
117.5–122.5	3
112.5–117.5	9
107.5–112.5	11
102.5–107.5	8
97.5–102.5	8
92.5– 97.5	4
87.5– 92.5	1
82.5– 87.5	2
77.5– 82.5	0
72.5– 77.5	1
67.5– 72.5	0
62.5– 67.5	0
57.5– 62.5	1
	$N = 71$

Step 1: $\frac{71}{2} = 35.5 =$ half-sum.

Step 2: $1 + 1 + 2 + 1 + 4 + 8 + 8 = 25$. Subtotal.

Step 3: $35.5 - 25 = 10.5$.

Step 4: $\frac{10.5}{11} = 0.954 \times 5$ (size of step) $= 4.77$.

Step 5: $107.5 + 4.77 = 112.27 =$ median.

Step 6: Check.

$35.5 - 35 = 0.5$.

$\frac{0.5}{11} \times 5 = 0.23$.

$112.5 - 0.23 = 112.27 =$ median.

perimental purposes has thus naturally tended to increase the popularity of the arithmetic mean as a measure of central tendency. In general, and in the absence of any other guiding principle, use the median in all interpretations or comparisons in which the median itself was used in securing the basis for comparison. That is, if test results are to be compared with test norms which are based upon medians, then the medians of the test results should be computed and used. Comparative norms based upon means may well be compared with class means. For most experimental purposes the arithmetic means should be used, particularly where the scores of individuals are compared with their own scores obtained under experimental controls. In most experimental studies it is desirable for all measures to receive consideration, and furthermore, in most such studies other measures, as the standard deviation, are required. Since these additional statistical measures are usually based upon the same processes as those used in calculating the mean, the arithmetic mean is the logical measure to use.

III. MEASURES OF VARIABILITY

The need for measures of variability in the interpretation of test scores arises through the fact that two groups of pupils may earn scores on a test which will have the same medians or arithmetic means and yet represent distinctly different types of instructional situations. At least two types of descriptive measures of a distribution are needed before all its essential features can be revealed. The measures of central tendency reveal the points on the scale where the typical scores are most likely to be found. Some method of expressing variability is required to reveal differences in range of talent.

The two groups of scores presented as Class A and Class B in Table 42 illustrate this situation very clearly. The means of the two

TABLE 42

ILLUSTRATION OF NEED FOR MEASURES OF VARIABILITY

Class A		Class B
122		98
116		96
108		95
101		93
96		90
92		89
89		87
86	Means	86
83		85
80		83
76		82
71		79
64		77
56		76
50		74

series of scores are identical, each being 86. The *range* of the scores for Class A is 72 (122 − 50), which is exactly three times the range (98 − 74 = 24) of the Class B scores. The *quartile deviation* computed from the ungrouped scores is 15 for Class A and 7 for Class B. The *standard deviations* of the scores are 20.16 and 7.3 for the Class A and Class B scores respectively. Even the most inexperienced teacher or student must recognize that very different ranges of ability are present in these two classes and that correspondingly different instructional problems are presented to the teacher.

110. The Range.

Of the three commonly used measures of variability mentioned in the illustration in the previous paragraph the *range* is the easiest to find and the least useful measure. *The range is the scale distance between the lowest and the highest scores in an array.* The very definition of the range makes it apparent that it is one of the least reliable measures, since it is so readily affected by the fluctuation of either of the extreme scores. In arrays of test scores or frequency tables in which the scores fall into line quite regularly the range may be a fairly consistent measure. In the illustration given in Table 30 the range is almost as effective in revealing the spread of ability as the standard deviation, which is usually considered to be one of the most reliable measures of dispersion. This, without doubt, may be traced to the consistency with which the test scores vary above and below the means. It may be sufficient to point out here that the range is rarely used as an evidence of dispersion or variability in the interpretation of educational-test scores since such scores rarely fit into the scale with the consistency shown in the illustration in Table 30.

111. Quartile Deviation.

The particular merit of the *semi-interquartile range* or *quartile deviation* (Q) as a measure of the variability of test scores lies in the fact that it utilizes the range of the middle half of the cases rather than the range of the extremes. In actual practice the range of the middle half of the cases is found by counting off frequencies from the lower end of the distribution until a point cutting off 25 per cent of the cases is located. The method of finding this point is identical with the procedure in computing the median except that only 25 per cent instead of 50 per cent of the cases in the distribution are considered. This point is commonly designated as Q_1. A point on the scale which cuts off 25 per cent of the cases from the top of the distribution is found in a similar way. This is known as Q_3. The remaining cases included between these two points are the middle 50 per cent. The reliability of this measure lies therefore in the fact that it is based upon the portion of the distribution in which the density of the population is greatest.

The quartile deviation (Q) is found by taking one-half of the difference in the scale values of the points Q_3 and Q_1. The formula for this measure of variability is $Q = \frac{Q_3 - Q_1}{2}$. The computation of Q is illustrated in terms of the data presented in Table 41, and since the

procedures are essentially the same as those used in finding the median the steps are summarized very briefly.

Step 1. Find 25 per cent of the cases. In this problem one-fourth of the cases equals 17.75.

Step 2. Summate the frequencies beginning at the bottom until a point not in excess of 17.75 cases is reached. $1 + 1 + 2 + 1 + 4 + 8 = 17$. The difference between this subtotal and 17.75 is 0.75 case.

Step 3. 0.75/8 times 5 equals 0.47 unit.

Step 4. Add 0.47 unit to the beginning of the interval in which the 8 cases are found. Thus, $102.5 + 0.47$ equals 102.67, which is Q_1 for this distribution.

Step 5. Summate the frequencies beginning at the top of the distribution. $1 + 1 + 3 + 2 + 5 + 4 = 16$. The difference between $N/4$ or 17.75 is 1.75 cases.

Step 6. 1.75/2 times 5 equals 4.38 units.

Step 7. Since this computation of Q_3 is proceeding from the *top* of the distribution the 4.38 units must be *subtracted* from the *top of the step* into which the interpolation is made. Thus, $132.5 - 4.38 = 128.12$, which is Q_3 for this distribution.

Step 8. $Q_3 - Q_1$ or $128.12 - 102.67$ equals 25.45, which is twice the value of the semi-interquartile range. Thus 25.45 divided by 2 equals 12.73, or the quartile deviation for this distribution.

The values Q and the median (Q_2) are frequently confused. They are quite different measures, however. The median or fiftieth percentile is a measure of central tendency; Q is a measure of the variation. Q expresses the variability of an array in terms of the average distance from the center of the distribution it is necessary to go in either direction to include the middle half of the cases.

112. Standard Deviation.

Such simple devices as the range and quartile deviation (Q) are sufficiently exact for many ordinary situations involving the interpretation of test results. However, other statistical analyses demand more refined measures of variability. For this type of work the *standard deviation* is generally used. *The standard deviation is the square root of the mean of the square of the deviations from the mean of a distribution.* Expressed in symbols the standard deviation is $S\sqrt{\frac{\Sigma FD^2}{N} - C^2}$ in which ΣFD^2 equals the deviations expressed in the form of the sum of the products of the frequencies at each step by the deviation of each step from the assumed mean; N equals the number of cases in the distribution; c equals the correction used in computing

the arithmetic mean; and s represents the size of the class interval of the distribution in units.

The standard deviation may be computed about any common measure of central tendency, although in common practice it is usually computed about the arithmetic mean. There is at least a theoretical advantage in using the mean as the point around which to determine the standard deviation. The arithmetic mean is the point in a distribution about which the deviations are the least.

113. Meaning of the Standard Deviation.

The likenesses and differences of the quartile deviation (Q) and the standard deviation (σ) are shown in Fig. 21. The quartile deviation, or semi-interquartile range, by definition takes into account the middle

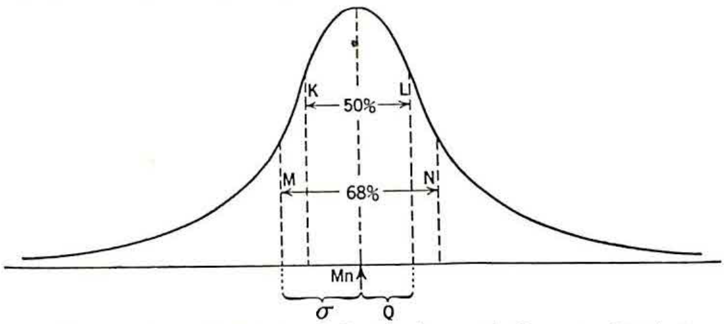


FIG. 21.—Comparison of Standard Deviation and Quarter Deviation (Q).

50 per cent of the cases. That is, the ordinates (lines erected perpendicular to the base line of the curve) erected at a scale distance equal to Q on either side of the mean or median include 50 per cent of the area of the surface between the curve and the base line. In the diagram the lines K and L represent the ordinates erected at a distance equal to Q on either side of the mean. The lines M and N represent ordinates erected at a distance equal to σ on either side of the mean. The standard deviation (σ) takes into account approximately 68 per cent (in a normal distribution 68.26 per cent) of the area of such a distribution.

In a normal distribution the value *sigma* bears a definite relationship to the curve of the distribution itself. When a normal distribution is presented in graphic form the result is a symmetrical bell-shaped curve with many cases in the middle and few at the extremes. Certain types of these characteristic bell-shaped distributions have come to be called *normal* curves. For these normal curves, formulas have been derived from which such typical curves may be computed if certain basic data concerning the curve are given. In these formu-

las *sigma* is one of the values which must be given in order to construct such a curve. Sigma, in the typical formula, represents the distance from the mean at which the curve changes from convex to concave. In Fig. 21 the points where the curve changes its character are indicated by the ordinates lettered M and N.

Thus, because of this direct mathematical relationship which the standard deviation bears to the curve of the distribution itself, and the reliable expression of variability which it provides due to the fact that every deviation in the distribution is considered, the standard deviation is one of the most useful of the measures of variability.

114. Computing the Standard Deviation (σ) from Ungrouped Data. In the computation of the standard deviation from ungrouped data, as in the accompanying illustration, the mean of the distribution must be found. When the data are grouped in a frequency table it is not strictly necessary for the arithmetic mean to be computed, although it is necessary to go through all but the last step of the process.

The steps in the computation of the standard deviation from ungrouped data are given in detail in connection with data from Table 42. See Table 43.

TABLE 43

DATA FOR CLASS A FROM TABLE 42

Test Scores	d (Deviation)	d^2 (Deviations Squared)
122	+36	1296
116	+30	900
108	+22	484
101	+15	225
96	+10	100
92	+6	36
89	+3	9
86	0	0
83	−3	9
80	−6	36
76	−10	100
71	−15	225
64	−22	484
56	−30	900
50	−36	1296
Total 1290		$\Sigma d^2 = 6100$
Mean 86		

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}}$$

$$= \sqrt{\frac{6100}{15}}$$

$$= \sqrt{406.67}$$

$$= 20.17$$

The mean of the scores for Class A in Table 43 is 86. Thus a score of 89 deviates from this mean by 3 points. A score of 96 deviates 10 points, etc. The standard deviation (σ) is the mean of the squares of these deviates from the mean of the array of scores. It is necessary therefore to square each of these deviates. These are given under the column headed d^2. Since each deviation appears only once and the data are ungrouped, the formula may be simplified to read $\sigma = \sqrt{\Sigma d^2/N}$. The sum of the deviations squared (Σd^2) is 6100. The mean of these deviations is therefore 406.67. This value is the mean of the deviations squared. Therefore to turn it into units of the scale the square root of this quantity must be taken. This value is found to be 20.17, which is the standard deviation (σ) of this series of scores. The mean of this distribution is 86. The σ is 20.17. This means that, between scores 20.17 points larger and 20.17 points smaller than this mean, approximately two-thirds (68.26 per cent) of the scores will be found.

115. Computing the Standard Deviation from Grouped Data.

The method of computing the standard deviation from ungrouped data illustrated in Table 43 may be applied with very few changes to the calculation of sigma from grouped data. A slight change in the general formula is required, for, when the scores are grouped in class intervals, the deviations of the scores must be considered by groups having the mid-points of the steps in which they are found. This permits the expression of the deviations in steps rather than in units of the scale. The formula for use in calculating the standard deviation when the data are grouped in a frequency distribution is $s\sqrt{\frac{\Sigma fd^2}{N} - c^2}$. The steps in the application of this formula in the calculation of the standard deviation of the scores originally presented in Table 38 will make clear all the processes involved in finding the sigma of a frequency distribution. The computations themselves are shown in Table 44.

Step 1. Assume a mean as near as possible to the true mean of the distribution in order that the correction (c) may be as small as possible. If the correction is larger than 0.5 step it may be desirable to start the work over and assume a mean nearer the true mean. In Table 44 the step with a mid-point of 7 is taken as the zero step.

Step 2. Lay off deviations above and below the step assumed as zero, and multiply the frequencies in each step by the deviation of the step exactly as in the calculation of the arithmetic mean.

Step 3. Summate plus fd's and the minus fd's algebraically. The

TABLE 44

DATA FROM TABLE 38

Class Intervals	Mid-Points	f	d	fd	fd^2
14.5–15.5	15	1	+8	8	64
13.5–14.5	14	0	+7	0	
12.5–13.5	13	0	+6	0	
11.5–12.5	12	1	+5	5	25
10.5–11.5	11	1	+4	4	16
9.5–10.5	10	2	+3	6	18
8.5– 9.5	9	2	+2	4	8
7.5– 8.5	8	3	+1	3	3
6.5– 7.5	7	4	0	+30	
5.5– 6.5	6	6	−1	6	6
4.5– 5.5	5	3	−2	6	12
3.5– 4.5	4	1	−3	3	9
2.5– 3.5	3	2	−4	8	32
1.5– 2.5	2	1	−5	5	25
		$N = 27$		−28	218

$$c = \frac{30 - 28}{27} = \frac{+2}{27} = +0.074.$$

$$c = 0.005.$$

$$\frac{\Sigma fd^2}{N} = \frac{218}{27} = 8.037.$$

$$8.037 - 0.005 = 8.032.$$

$$\sigma_{\text{steps}} = 8.032 = 2.83.$$

Step of 1 in table, therefore $\sigma = 2.83$.

sum of the fd's in this problem is + 2 units. Divide this by the number of cases in the table ($N = 27$), and the resulting correction (c) is + 0.074. This correction is the same as the one used in computing the mean.

Step 4. Square this correction in order to have it in the same denomination as the values from which it must later be subtracted. The square of c (+ 0.074) is 0.005 in this problem.

Step 5. Multiply the values under the column headed fd by the values under d. This will give a column of values known as fd^2. Summate this column. In this problem the sum of the fd^2 is 218.

Step 6. Divide the sum of the fd^2 by N, the number of cases in the distribution. The result of this division is 8.037.

Step 7. Since this value, 8.037, is always too large in proportion to the amount the true mean deviates from the assumed mean (in this case, the amount represented by the value c) it must be reduced an amount equal to the square of c. Thus,

$$8.037 - 0.005 = 8.032.$$

Step 8. The sigma of this distribution expressed in steps is next obtained by extracting the square root of the value 8.032. The square root of this value to two decimal places is 2.83. Since the class intervals used in this frequency table are steps of one unit, the standard deviation is therefore 2.83.[5]

116. Using the Standard Deviation.

Assignment of Class Grades. The student or teacher who is interested in the critical analysis of test scores will find the standard deviation a very useful and reliable instrument for the purpose. For example, it offers the basis for a practical plan for turning scores on objective tests into class marks. The importance of this practice is so great that the steps involved in the technique are given in detail. The computations described are based upon the objective test scores from a class of 45 pupils given in Table 45. The student will do well to check all these computations for errors.

Step 1. Prepare a suitable frequency table of the test scores, lay off the deviations from the assumed mean, and find the sum of the fd's and the arithmetic mean. The mean of this distribution is 68.55.

Step 2. Compute the standard deviation (σ) of this distribution by multiplying the fd's by the deviations in steps, summating the fd^2's, dividing this sum of the fd^2's by the number of cases, subtracting from this quotient the square of the c used in finding the arithmetic mean, extracting the square root of this remainder, and multiplying this root by the size of the step used in the table. The standard deviation of this distribution found in this manner is 19.40.

Step 3. Since a distance of two and one-half sigma units above and below the mean includes almost 99 per cent of all cases in a distribution this number of sigma units is laid off above and below the mean. This naturally results in placing one of the sigma units in the middle of the distribution in such a way that one-half of the sigma distance of the middle unit extends above the mean and one-half

[5] Problems 19, 20, and 21 in Greene's *Work-Book in Educational Measurements* (Longmans) furnish excellent supplementary practice in the computation of the standard deviation.

TABLE 45

STANDARD DEVIATION TECHNIQUE FOR ASSIGNING CLASS GRADES *

Test Scores		Mid-Points	Class Intervals	f	d	fd	fd^2
109		110	107.5–112.5	1	10	10	100
104		105	102.5–107.5	2	9	18	162
103	A(11.1%)	100	97.5–102.5	2	8	16	128
102		95	92.5– 97.5	4	7	28	196
99		90	87.5– 92.5	0	6	0	0
95		85	82.5– 87.5	2	5	10	50
95		80	77.5– 82.5	2	4	8	32
94		75	72.5– 77.5	3	3	9	27
93		70	67.5– 72.5	3	2	6	12
84	B(17.8%)	65	62.5– 67.5	4	1	4(+ 109)	4
83		60	57.5– 62.5	7			
79		55	52.5– 57.5	7	−1	−7	7
79		50	47.5– 52.5	4	−2	−8	16
77		45	42.5– 47.5	1	−3	−3	9
76		40	37.5– 42.5	1	−4	−4	16
76		35	32.5– 37.5	2	−5	−10(− 32)	50
71				$N = 45$		77	809
71							
69							
64							
64	C(35.5%)						
64							
64							
62							
60							
60							
60							
59							
59							
58							
57							
57							
56							
56							
55							
55							
53	D(31.1%)						
52							
52							
51							
51							
47							
41							
37	Fd(4.5%)						
37							

$$\text{A.M.} = 60 + s\frac{\Sigma fd}{N} = 60 + 5\cdot\frac{77}{45} = 60 + 5(1.71). = 60 + 8.55. = 68.55.$$

$$\text{S.D.} = s\sqrt{\frac{\Sigma fd^2}{N} - c^2}. = 5\sqrt{\frac{809}{45} - (1.71)^2} = 5\sqrt{17.98 - 2.92.} = 5\sqrt{15.06.} = 5 \times 3.88 \text{ or } 19.40.$$

Find score limits:

$68.55 + \frac{1}{2}(19.40) = 78.25$ upper limit of C group.

$68.55 + 1\frac{1}{2}(19.40) = 97.65$ upper limit of B group.

$68.55 - \frac{1}{2}(19.40) = 58.85$ upper limit of D group.

$68.55 - 1\frac{1}{2}(19.40) = 39.45$ upper limit of F group.

A = above 97.65.

B = 78.25 to 97.65.

C = 58.85 to 78.25.

D = 39.45 to 58.85.

Fd = below 39.45.

* For a complete explanation and discussion of the many problems involved in objectifying teachers marks see Bangs, C. W., and Greene, H. A., "Teachers' Marks and the Marking System," University of Iowa Extension Bulletin No. 244, College of Education Series No. 26, May 15, 1930.

below. Accordingly, to the arithmetic mean of 68.55 add one-half of the standard deviation (one-half of 19.40). This gives a value of 78.25, which becomes the upper limit of the group of scores which will be assigned grades of C.

Step 4. The upper limit of the group of scores to be assigned B's is found by adding one and one-half standard deviation units to the arithmetic mean. Thus, $68.55 + 1.5\ (19.40) = 97.65$, which is the upper limit of the B group.

Step 5. The upper limit of the D group is found by subtracting one-half of a standard deviation unit from the mean. $68.55 - 0.5\ (19.40) = 58.85$.

Step 6. The upper limit of the Fd group is obtained by subtracting one and one-half sigma units from the mean of the distribution. $68.55 - 1.5\ (19.40) = 39.45$.

Step 7. From these values the score limits of this distribution may be set up. Class grades may then be assigned as indicated to the scores within the limits specified.

Grades	Score Limits
A	97.65 and above
B	78.25 to 97.65
C	58.85 to 78.25
D	39.45 to 58.85
Fd	below 39.45

It is readily apparent that practically no subjective factors are involved in the assignment of grades by this method. The objective test scores of the 45 pupils used in the illustration are changed by this treatment into 5 A's, 8 B's, 16 C's, 14 D's, and 2 Fd's. The score limits are determined by the standard deviation units and would be the same no matter who assigned the grades. It should be noted, however, that *these limits hold only for this particular distribution and must not be assumed to be true for any other test.* The teacher should also remember that this method of grading does not take into account the absolute level of ability at which a particular class works. The superior pupil in an average or poor class receives an A by this method just as readily as does the superior pupil in a very superior class. This is probably less serious than it sounds, however, for most class groups large enough to warrant the application of this technique average out quite well in this respect.

Basis for T-Scores and Other Derived Scores. The standard deviation also affords the basis for derivation of a number of useful

derived scores in test interpretation. For example, the T-score now commonly used in commensurating test scores depends upon the standard deviation for its basic unit. For many years prior to the development and popularization of the T-score, test scores were expressed in terms of their position in the total distribution. For instance, a pupil's score might be a member of a distribution having a mean of 60 points and standard deviation of 5 points. A score of 50 in the test would be designated by a standard score of -2.0 sigma units, since it lies exactly two standard deviation units below the mean of the distribution. This same procedure is used in assigning T-scores. The formula for the T-score is $T = \frac{(m - x)\ 10}{\sigma} + 50$, in which m is the mean of the distribution, x the deviation of the score, and σ the standard deviation of the distribution. The difference between the mean and any test score is multiplied by 10 in order to remove all decimal points. The 50 points are added in order that there may be no negative scores. The T-score is a very convenient method of interpreting test scores. T-scores of 25, 50, and 75 mean that the individual pupil's scores were right at the lower quartile, the median, and the upper quartiles. This fact makes it easy to attach meaning to the test scores.

Scaling of Test Items. The standard deviation, along with certain other measures of variability, represents a convenient unit in which to evaluate the difficulty of test items. When used under these conditions the standard deviation of a theoretically normal curve of the specified ability is used as the unit in laying off differences in difficulty along a linear scale. As a first step in the procedure, the percentage of pupils failing on each item or exercise must be secured. By means of tables based upon the normal curve these percentages of failure are changed into standard deviation units which express the position of the exercises with respect to the mean ability of an infinite and normal population. Exercises which are answered successfully by 50 per cent of the class are assigned a position at the mean. Exercises missed by 55 or 60 per cent of the class are given sigma values above the mean, etc. A significant feature of this procedure, however, is the fact that a difference in difficulty of 5 per cent near the mean results in a relatively small sigma difference, while a 5 per cent difference near the extremes of the distribution makes a relatively large sigma difference. This is in conformity with the fact that because of the height of the curve near the mean a smaller distance along the linear scale on the base line is required to add a given area of the curve. Thus, the difference in the sigma values assigned to two

test items having percentages of failure of 55 and 60 is 0.13 standard deviation unit (2.74-2.61),[6] while the difference in apparent difficulty of two items failed by 90 and 95 per cent of an experimental group is 0.34 standard deviation unit (4.09–3.75). The net result of this method of item evaluation is to magnify somewhat the simplicity of the very easy items and the difficulty of the very hard ones.

Sigma units are also utilized in the construction of scales for the estimation of the merit or quality of certain classroom products. The use of these units in the derivation of such scales is discussed in detail in Chapter XII of this book.

IV. MEASURES OF RELATIONSHIP

The critical analysis and interpretation of educational test results often make it necessary for the teacher and research student to secure more exacting descriptions of the situation than are afforded by the simple measures of central tendency and variability. For example, the matter of the selection of the test itself is one which cannot be determined wholly on the basis of the arithmetic means and the standard deviations. The most useful information for this purpose is found by determining the relationship which exists between the ability to be measured and the tests or measures under consideration. In the construction and use of informal objective examinations there are occasions when it is necessary to discover exactly how accurately the examination measures, and how much this accuracy would be improved by increasing the length of the examination. This type of analysis also requires the use of the method of correlation, the method which permits the determination of relationships among different measures of the same individuals.

117. The Correlation Coefficient.

In the statistical expression of relationships, as in the other measures, it is desirable that this relation between two series of variables be expressed in a single objective or mathematical value. A number of different ways have been proposed for the derivation of these expressions of relationship, but no one of them has produced a term which is both objective and easily interpreted. Methods of computing relationships in terms of the correspondence between rank positions of scores, and in terms of the percentage of the scores falling within a specified unit of variability of each other, have been pro-

[6] See Table 5 on page 392 of Rugg's *Statistical Methods Applied to Education* (Houghton Miffln), or any similar table of sigma values.

posed. In general these procedures lack sufficient exactness to warrant their common use in the analysis of test results. The student who is interested in these different methods of revealing relationships will find them discussed in certain of the treatments on statistical methods listed in the references at the end of this chapter. In this discussion, attention is given exclusively to the Pearson product-moment method, which is not only the basic method but also the one most commonly used in educational statistics.

The index expressing the degree of relationship between two series of measures is called a *coefficient of correlation.* The coefficient resulting from the application of the Pearson product-moment method is designated as r. The possible values of r range from perfect positive relationships ($+ 1.0$) through all the possible decimal values through zero to $- 1.0$ indicating a perfect negative relationship. An r of zero is taken to mean that no relationship exists between the measures or that it is entirely due to chance. Positive relationships may be expected between such factors as barometric readings and atmospheric pressure, or between each of such factors as native capacity, initiative, effort, concentration, interest, and school accomplishment in a given field. Negative relationships are usually found to exist within a given school grade between the chronological age of the pupils and their achievement scores on a reliable achievement test. Many low or zero relationships are found in educational data, but the best illustration of this type of correlation is one in which pure chance operates. If two packs of numbered cards are shuffled carefully and cards are drawn from each pack and paired, the resulting relationship is due to pure chance, and the coefficient of correlation (r) approaches zero. If the same packs of cards are rearranged so that the numbers appear in ascending order in each pack and cards are drawn off the top of each and paired as before, the r obtained should be positive and very high. If one of the packs is inverted and the cards are drawn as before, the result should be a high negative correlation.

The Pearson product-moment method of computing correlations, while involving a large number of rather complicated calculations, really calls for very few skills that the student has not encountered previously in this work. This coefficient of correlation (r) is a single numerical value which expresses the tendency of corresponding pairs of measures of two fields to deviate similarly from their respective means. Modern methods of work permit the computation of this coefficient from data arranged in frequency tables of the double-entry type.

The double-entry or correlation table permits the simultaneous

tabulation of two measures of the same individuals. The class intervals are set up exactly the same as in preparing a simple frequency table. In fact, it is merely an overlapping table with two sets of class intervals, one reading upward along the left-hand margin and one reading to the right along the top. Such a double-entry tabulation is shown in Table 46, which also serves to illustrate the specific steps involved in computing the correlation coefficient.

The data in Table 46 represent the very real problem of determining the reliability of an experimental test by finding the correlation of the scores made on one-half of the test with the scores made on the other half. Let us assume that Pupil A made scores of 25 and 29 on the two halves of the test. The position of the score of 25 on the first half is found in the scale for that part of the test. This is in the step with a mid-point of 24. We then move across the table horizontally until the proper space is found for the score of 29 on the second half of the test. This is in the step with a mid-point of 30. A single tabulation mark in that space identifies both scores and at the same time indicates something of their relation to each other. In such a table a tendency for the frequencies to group themselves along the diagonal of the table itself is an indication of a positive relationship. Scores which deviate from the diagonal reduce the relationships. Figure 22 indicates something of the types of relationship which may

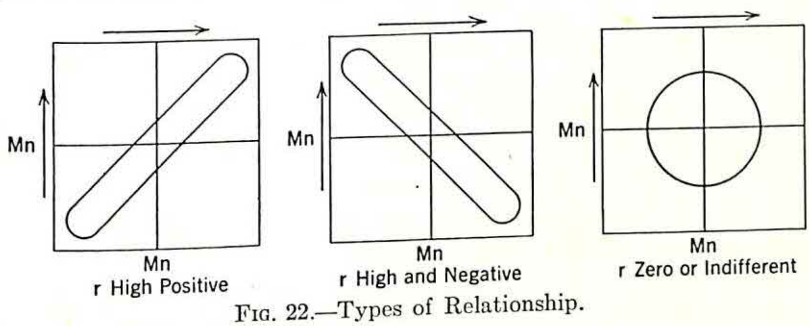


FIG. 22.—Types of Relationship.

be expected from characteristic groupings of data. This interpretation can be made only when the scales of the tables read upward and to the right.

118. Computation of Pearson Coefficient of Correlation.

The Pearson product-moment coefficient of correlation is found by the solution of the following formula:

$$r = \frac{\frac{\Sigma xy}{N} - c_x c_y}{\sigma_x \sigma_y}$$

in which N is the number of cases in the distribution, σ_x the standard deviation of the distribution on the x-axis, σ_y the standard deviation on the y-axis, c_x the correction on the x-axis, c_y the correction on the Σy-axis, and xy the summation of the products of the deviations of each measure from the means of the distributions.

The following steps are involved in the solution of the formula as applied to the data given in Table 46:

Step 1. The data are already tabulated in a double-entry or correlation table, so the first step in the work is to total the frequencies on each axis and cross-check the totals.

Step 2. Assume suitable means for each axis of the table and lay off steps above and below the zero step. Compute the corrections on the x-axis and on the y-axis exactly as in finding the arithmetic mean and the standard deviation. In this work the c_x is $+ 0.307$ and the c_y is $- 0.109$.

Step 3. Multiply the fd column by the d column and summate the resulting fd^2's separately for each distribution. Complete the calculation of the standard deviation for each distribution. *Leave the resulting sigmas in terms of steps.* This will save an extra multiplication and unnecessary division later in the work.

Step 4. Find the sum of the xy products. This constitutes the only absolutely new step in the process up to this point. In the product-moment method each score or frequency is weighted in proportion to the distance it lies away from the means of the distributions. Thus, in the example, the single individual score which lies in the extreme upper right-hand section of the table deviates a distance of $+ 6$ steps from the mean of the y-axis and $+ 7$ steps from the mean of the x-axis. The combined moment of this single case is found by multiplying it by the product of 6 and 7 $(1 \times +6 \times +7 = 42)$. This case therefore has a moment of 42. The three cases at the intersection of the steps with mid-points of 33 have a combined product of 90 $(+5 \times \; +6 \times 3 = 90)$. All cases in the upper right-hand and lower left-hand quadrants have a positive moment, owing to the algebraic law of signs in multiplication. All frequencies lying in the upper left-hand or the lower right-hand quadrants of the table have negative product-moments since the product of a plus step-deviation by a minus step-deviation results in a negative sign. In this work the xy products are summated algebraically in a column along the right-hand side of the table. The total of the xy products is 1780.

Step 5. The numerator of the fractional equation representing the

TABLE 46

CORRELATION TABLE SHOWING RELATION OF HALF-HALF SCORES ON IOWA PLANE GEOMETRY APTITUDE TEST (1933)

Score on second half of test: → x

Score on first half of test: → y

Mid-points:	0	3	6	9	12	15	18	21	24	27	30	33	36	f	d	fd	fd^2	xy
36													1 (42)	1	6	6	36	42
33										2 (40)	4 (100)	3 (90)	1 (35)	10	5	50	250	265
30							2 (8)	2 (16)	5 (60)	8 (128)	2 (40)			19	4	76	304	252
27							2 (6)	6 (36)	11 (99)	3 (36)	2 (30)			24	3	72	216	207
24						2	12 (24)	8 (32)	15 (90)	7 (56)	1 (10)			45	2	90	180	212
21					2 (−2)	9	14 (14)	22 (44)	5 (15)	1 (4)				53	1	53	53	75
18				4	17	23	19	4	3					70	0	347		
15		1 (4)	1 (3)	13 (26)	19 (19)	24	13 (−13)	4 (−8)						75	−1	75	75	31
12		1 (8)	7 (42)	18 (72)	19 (38)	16								61	−2	122	244	160
9		2 (24)	14 (126)	11 (66)	8 (24)	1								36	−3	108	324	240
6		5 (80)	1 (12)	8 (64)										14	−4	56	224	156
3	2 (50)	3 (60)												5	−5	25	125	110
0	1 (30)													1	−6	6	36	30
																	2067	1780
f	3	12	23	54	65	75	62	46	39	21	9	3	2	414				
d	−5	−4	−3	−2	−1		1	2	3	4	5	6	7					
fd	15	48	69	108	65	305	62	92	117	84	45	18	14	+432				
fd^2	75	192	207	216	65		62	184	351	336	225	108	98	2119				

$$c_y = \frac{-392 - 347}{414} = -0.109.$$

$$c_y^2 = 0.01.$$

$$\frac{\Sigma fd^2}{N} = \frac{2067}{414} = 4.99.$$

$$4.99 - 0.01 = 4.98.$$

$$\sigma_y = \sqrt{4.98} = 2.23.$$

$$c_x = \frac{432 - 305}{N} = \frac{+127}{414} = +0.307.$$

$$c_x^2 = 0.09.$$

$$\frac{\Sigma fd^2}{N} = \frac{2119}{414} = 5.12. \quad 5.12 - 0.09 = 5.03.$$

$$\sigma_x = \sqrt{5.03} = 2.24.$$

$$r = \frac{\frac{\Sigma xy}{N} - c_x c_y}{\sigma_x \sigma_y} = \frac{\frac{+1780}{414} - -0.0335}{4.995}$$

$$= \frac{+4.33}{4.995} = +.867.$$

correlation coefficient is $\frac{\Sigma xy}{N} - c_xc_y$. This quantity is found by dividing 1780 by the number of cases in the distribution and subtracting (algebraically) the product of the corrections for the two distributions. The result of dividing 1780 by 414 is 4.2995. The c_xc_y product is -0.0335. Since this c_xc_y product is negative (owing to the negative sign of one of the corrections), the net result is the addition of these two quantities. The numerator of the fractional part of the formula now becomes 4.33.

Step 6. The denominator of the fractional part of the formula comprises the product of the two standard deviations. The $\sigma_x\sigma_y$ product for this correlation table is 4.995.

Step 7. The correlation coefficient (r) is the ratio of the two values found in steps 6 and 7. The r of this distribution is therefore $+.867$, which means that the relationship is positive and significantly high.

119. Meaning of the Correlation Coefficient.

It was suggested earlier in this discussion that the interpretation of the correlation coefficient is probably much more difficult than its computation. There are a few devices which are helpful to the inexperienced worker, but, in general, assurance in the interpretation of these measures comes only through experience. The suggestions given in this section may be useful during the period when this experience is being gained.

One of the important outcomes of the use of correlation methods is that within certain limits it makes possible the estimating of unknown values from known values. The accuracy of this estimate, however, depends directly upon the correlation between the factors measured. For example, if it is known from previous experience that there is a high positive relationship between achievement in a specific manual arts subject and the students' response to certain manual dexterity tests, the probable achievement of a group of prospective students in this manual arts course may be determined within limits by securing their response to the manual dexterity test. A correlation coefficient of $+1.0$ for the two factors would mean that an estimate of accomplishment based on the one factor would be approximately 100 per cent correct. As the amount of the correlation decreases the accuracy of the forecast declines, but not in a direct manner. A correlation of $+1.0$ may mean 100 per cent accuracy in the estimate based on the relationship; but a correlation of $+.50$ does not mean at all that the estimate based on it will be 50 per cent correct. A glance at the accompanying table will demonstrate this interesting fact about the correlation coefficient.

The percentages of forcasting accuracy for different values of r given in Table 47 are obtained by applying Kelley's proposed formula for the *coefficient of alienation* $(k = \sqrt{1 - r^2})$ and then deducting the resulting values expressed as percentages from 100 per cent. If estimates of one variable are to be made from measurements of another related variable, this table will prove to be a useful safeguard.

TABLE 47

PERCENTAGE OF FORECASTING ACCURACY FOR SPECIFIC VALUES OF r

Coefficient of Correlation	Percentage of Forecasting Efficiency
1.00	100
.99	86
.98	80
.95	69
.90	56
.866	50
.80	40
.75	34
.70	29
.65	24
.60	20
.50	13
.40	8
.30	5
.20	2
.10	½

The following illustrations and practical interpretations of typical correlation coefficients representative of the sort obtained from educational data have been gleaned from a number of sources. They are offered here for whatever guidance they may give to the student or teacher interested in this type of test analysis.

r value	Educational Situation	Interpretation
+ .96	Relation of two forms of a long, analytical reading test for high-school students.	Evidence of unusually high reliability of measurement; treat scores with confidence.
+ .80	Repetition of same form of a group test of mental ability after a lapse of one semester.	Evidence of a marked relationship; considerable prognostic power even after lapse of a long interval.
+ .65	A composite of three separate estimates by same teacher of the ability of a class of 35 students to respond to an objective test in industrial arts.	Evidence of some relationship but of limited use for prognostic purposes.
+ .50	Scores on a good group intelligence test and the class grades of a class in industrial arts.	A very slight relationship of no practical value for forecasting purposes (only 13 per cent effective).
— .24	Chronological ages of pupils in a given grade and achievement scores on an objective test.	Negative relationship of an indifferent sort. Merely shows a very slight tendency for younger pupils in a grade to achieve at a higher level than the average.

120. Practical Uses of Correlation Method.

The teacher of industrial arts or the student of measurement in this field will unquestionably find the greatest use for correlation techniques in connection with the analysis of objective tests. The validity of a test may be expressed statistically in terms of the correlation of the instrument with some other criterion, such as another measure of known validity. The determination of the reliability of a test is almost entirely a matter of correlation method. Mastery of these uses of the correlation techniques will make the teacher a better critic of commercial standardized materials as well as a more independent and efficient builder and critic of teacher-made tests for classroom use. Such mastery can come only from extensive practice on problems calling for the use of these skills.[7] Students who are interested in the theory of correlation or in the use of correlation methods in more critical research with tests are referred to the many excellent textbooks on statistical methods now available.

V. ASSIGNMENT OF RELATIVE AND ABSOLUTE RANKS

121. Relative Ranks.

Achievement as expressed in test scores can have meaning only when it is possible to consider it in relation to some other, known level of achievement. In many cases, as when informal objective examinations are used, no definite standards or norms of achievement are available. Some simple method of comparing the accomplishment of each pupil in relation to the other individuals in the class is essential. The process of assigning relative ranks to pupils' scores in terms of their size is one way of doing this. This is accomplished by assigning to the individual making the highest score the first position in the class, the pupil making the second highest score the second position, etc. The assignment of such relative ranks is quite simple where the individual pupils make different scores, or where no tied scores appear. The illustration given in Table 48 shows how all such tied scores are treated in the assignment of relative ranks.

TABLE 48

Pupil	Score	Rank
A	15	1
B	12	2
C	11	3
D	10	4.5
E	10	4.5
F	9	7
G	9	7
H	9	7
I	8	9
J	7	10

Pupil A, with a score of 15 points, is assigned first position. Pupil B, with a score of 12, is given second position. Pupil C, with a score of 11, is given third place. Pupils D and E, both with scores of 10,

[7] See Problems 22, 23, 24, 25, 39, and 40 in Greene's *Work-Book in Educational Measurements* (Longmans).

would normally be assigned fourth or fifth places, but since it is impossible to assign fourth or fifth place to one rather than the other, tied or average rank is assigned to each. In this instance 4.5 position is given to each of the pupils making a score of 10 points. Pupils F, G, and H are also tied with scores of 9 points each, but since they would regularly be assigned sixth, seventh, and eighth positions they are each given the average of these positions, or seventh place. The practice, therefore, in all cases of tied scores is to assign average rank to the tied scores. When the number of tied scores is even, the ranks assigned will lie mid-way between the ranks which would ordinarily be assigned to two middle scores. When the number of tied scores is odd, the position assigned to all is the position which would normally fall to the middle score. In general, the position assigned to the pupil with the lowest score will agree with the number of cases in the series except when the last scores are tied.

122. Absolute Ranks.

The practice of assigning relative positions to pupils on the basis of their test scores, though aiding in the interpretation in some ways, actually covers up something of the actual situation. As a matter of fact, the assignment of relative ranks covers up the true differences in the size of scores. In Table 48 the difference of three score points between Pupils A and B is indicated by only a single position in rank the same as is given to the difference of one point for Pupils B and C. Thus, relative ranks reveal that a pupil is above or below another in achievement, but they do not indicate in any way the magnitude of that difference. Relative ranks also take no account of the actual achievement level at which the accomplishment takes place. A pupil having a rank of 18 in a class of 20 would be considered as having a low ranking in his group. However, if he were found to rank eighteenth among 400 similar individuals his position would indicate a significantly different type of achievement. Percentile ranks, as one form of absolute ranking, take this factor into account by reducing all ranks to a basis of 100 units. A percentile rank of 100 means that the individual making the specified score achieves at a level high enough to exceed 100 per cent of a similar group without regard to the number in the group. In a similar way, a percentile score of 75 means that the individual made a score such that it exceeds that of 75 per cent of the individuals of his group without respect to number.

Percentile scores are easily computed from frequency tables and are very useful in comparing the achievement of pupils taking an informal or non-standardized test. Percentile scores are also used very widely in the interpretation of standard test scores at the secondary-

school and college level. The student will recognize the seventy-fifth percentile as a measure with which he has already had some contact. This percentile is the same as the third or upper quartile (Q_3). It is found by exactly the same methods as are used in finding the median except that 75 per cent of the cases in the distribution are counted out below the point on the scale assigned as the seventy-fifth percentile. The same general methods are applied in the determination of the twenty-fifth, fiftieth, or any other designated percentile. The tenth, twentieth, thirtieth, and fortieth percentiles, etc., are known as the *deciles*. These are very often used in test interpretation.

The computation of the commonly used percentile scores is illustrated in Table 49. Since all the processes involved here have been used in earlier work, the computation is presented without comment.

TABLE 49

COMPUTATION OF PERCENTILE SCORES

Class Interval	f
162.5–167.5	1
157.5–162.5	1
152.5–157.5	0
147.5–152.5	3
142.5–147.5	2
137.5–142.5	5
132.5–137.5	4
127.5–132.5	2
122.5–127.5	5
117.5–122.5	3
112.5–117.5	9
107.5–112.5	11
102.5–107.5	8
97.5–102.5	8
92.5– 97.5	4
87.5– 92.5	1
82.5– 87.5	2
77.5– 82.5	0
72.5– 77.5	1
67.5– 72.5	0
62.5– 67.5	0
57.5– 62.5	1
	$N=71$

Percentile Score	Interpretation	Test Score
100	Score equaled or excelled by no student.	167
90	Score equaled or excelled by 10 per cent of students.	142
80	Score equaled or excelled by 20 per cent of students.	135
75	Third quartile—score equaled or excelled by 25 per cent of students.	128
70	Score equaled or excelled by 30 per cent of students.	124
60	Score equaled or excelled by 40 per cent of students.	116
50	Median—score equaled or excelled by 50 per cent of students.	112
40	Score equaled or excelled by 60 per cent of students.	109
30	Score equaled or excelled by 70 per cent of students.	105
25	First quartile—score equaled or excelled by 75 per cent of students.	103
20	Score equaled or excelled by 80 per cent of students.	101
10	Score equaled or excelled by 90 per cent of students.	95
0	Score equaled or excelled by practically all students.	58

The interpretation of percentile scores frequently gives some trouble to the worker inexperienced in their use. Fig. 23 is a graphic presentation of the percentile scores given in Table 49. This figure shows the characteristic curve (ogive) resulting from the use of per-

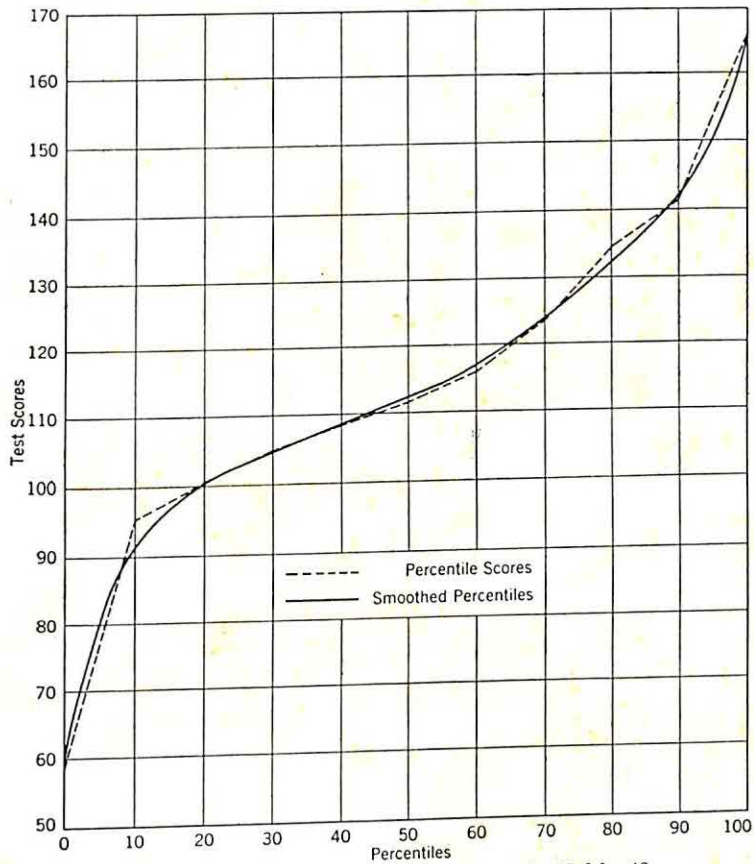


FIG. 23.—Ogive Curve Based on Data in Table 49.

centile scores. The heavy solid line in the figures represents the results of an arbitrary smoothing of these percentile scores. This smoothing process is frequently used when percentile scores are based on fairly large populations and are set up as tentative norms for the interpretation of the tests.

VI. SUMMARY

This chapter presents a non-technical discussion of a few of the common statistical tools which teachers of industrial arts will find useful in the analysis and interpretation of educational test results. Discussions of four of the six major statistical techniques outlined in

the introductory paragraph of this chapter are presented. The fundamental principles of the grouping and tabulating of test scores are stated and illustrated. The need for measures of dispersion is shown. The quartile deviation and the standard deviation are explained in some detail, and practical applications of these measures to problems of test analysis are made. The general meaning and the methods of correlation are given, along with a few definite hints concerning the interpretation of correlation coefficients. The practical uses and meanings of the ranking of test scores on both the relative and the absolute basis are discussed. The two remaining problems, dealing with the derivation and interpretation of test norms and standards, and the use of simple graphic methods of presenting the results of statistical analysis, are reserved for treatment in the following chapter.

There has been no attempt to make this chapter a complete discussion of all the interesting or even useful statistical techniques. To do this would require a volume in itself. As a matter of fact, the brevity of the treatment makes it impossible to present an adequate number of examples and illustrations to give the inexperienced worker sufficient experience with statistical problems. Real mastery of these skills can come only through repeated and continuous use. The student who is interested in achieving real skill and understanding in this field will wish to make extensive use of the selected references on page 224.

EXERCISES IN SUMMARIZING RESULTS OF TESTING

Tabulating Test Scores

Problem 1

a. Arrange or rank these scores from an objective examination in woodworking in descending order:

   95, 99, 40, 44, 68, 84, 54, 60, 91, 58, 66, 66, 72, 87, 77, 76, 65, 66, 70, 89, 77, 80, 78, 78, 62, 64, 64, 64, 54, 57, 58, 63, 93, 90, 76, 59, 62, 69, 70, 85, 72, 73, 83, 71

b. What is the largest score made on this test?

c. What is the smallest score made on this test?

d. What is the range of the scores?

e. If a frequency table with a step of 3 is made, how many steps will be required?

f. What will be the limits of the step required for the largest score?

g. What will be the mid-point of this step?

h. Make a frequency table of these scores using a 3-point step and mid-points divisible by the size of the step. Do your work on the left half of a sheet of paper and preserve it for use in later problem work.

i. If your work is right, the frequencies reading from the top will be as follows:

   1, 1, 2, 2, 1, 3, 1, 4, 2, 3, 5, 4, 6, 2, 3, 2, 0, 0, 1, 0, 1 $N = 44$

Computing the Arithmetic Mean

Problem 2

Compute the arithmetic mean from the frequency table prepared in Problem 1. (Answer = 70.4 from frequency table with step of 3.)

Computing the Mid-Measure and the Median

Problem 3

Compute the median from the frequency table in Problem 1.
(Answer = 69.7.)

Problem 4

Find the mid-measure for the scores given in Part *a* of Problem 1.
(Answer = 70.0.)

Computing Measures of Variability

Problem 5

Find the quartile deviation for the scores tabulated in Problem 1.
(Answer = 8.5.)

Problem 6

Find the standard deviation of the scores in the table prepared for Problem 1.
(Answer = 13.3.)

Computing Measures of Relationship

Problem 7

The following paired scores were obtained by giving the same form of an objective examination two times to the same pupils:

Pupil	1st Test	2nd Test	Pupil	1st Test	2nd Test
A	61	67	N	52	55
B	56	60	O	44	49
C	73	76	P	40	41
D	67	70	Q	33	33
E	53	49	R	58	59
F	48	52	S	76	77
G	43	44	T	70	75
H	35	31	U	63	63
I	23	25	V	50	50
J	57	56	W	41	46
K	78	81	X	36	37
L	71	73	Y	34	31
M	65	67	Z	25	27

a. Prepare a correlation table of these 26 pairs of scores. Use a 3-point step on both axes. Compute the coefficient of correlation as a basis for expressing the reliability of the objective examination. (Answer = +.959.)

Computing Percentile Ranks

Problem 8

Use the frequency table for the first test scores tabulated in Problem 7, and compute the percentile scores for each of the deciles as shown in Table 49. Check your own work for accuracy.

SELECTED REFERENCES

Garrett, H. E., *Statistics in Psychology and Education.* New York: Longmans, Green and Company, 1926.

Good, Warren R., *The Elements of Statistics.* Ann Arbor: The Ann Arbor Press, 1933.

Greene, H. A., *Work-Book in Educational Measurements.* New York: Longmans, Green and Company, 1930.

Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests.* New York: Longmans, Green and Company, 1935.

Gregory, C. A., *Fundamentals of Educational Measurement.* New York: D. Appleton Company, 1922.

Holzinger, Karl J., *Statistical Methods for Students in Education.* Boston: Ginn and Company, 1928.

Holzinger, Karl J., *Statistical Tables for Students in Education and Psychology.* Chicago: University of Chicago Press, 1925.

Holzinger, Karl J., and Mitchell, B. C., *Exercise Manual in Statistics.* Boston: Ginn and Company, 1929.

Kelley, T. L., *Statistical Method.* New York: The Macmillan Company, 1923.

Kramer, Edna E., *Educational Statistics.* New York: John Wiley and Sons, Inc., 1935.

Lindquist, E. F., and Stoddard, George D., *Study Manual in Elementary Statistics.* New York: Longmans, Green and Company, 1929.

Odell, C. W., *Educational Statistics.* New York: The Century Company, 1925.

Otis, Arthur S., *Statistical Method in Educational Measurement.* Yonkers, New York: World Book Company, 1925.

Rugg, Harold O., *A Primer of Graphics and Statistics for Teachers.* Boston: Houghton Mifflin Company, 1925.

Trabue, M. R., *Measuring Results in Education.* New York: American Book Company, 1924.

CHAPTER XV

INTERPRETING THE RESULTS OF TESTING

I. THE RESULTS OF TESTING

123. The Meaning of a Test Score.

It is important in this chapter on the interpretation of the results of testing to define clearly what is meant by a *test score*. In order to accomplish this, two or three new concepts may require explanation. In the first place, *a test score is a numerical expression of performance* on the part of an individual. Sometimes the test score is merely the number of exercises responded to correctly. Sometimes it is an arbitrarily defined scale value. But whatever its form, its function is to reveal in a quantitative way the performance of an individual as he responds to stimuli given under certain conditions. This leads to the second concept involved in the meaning of a score. The test score is an evidence of performance. Performance, the response of the individual to the test situation, is the expression of ability operating under certain conditions. The pupil may make a poor score because he does not have the ability to do better—may not know the facts. On the other hand, he may make a low score because of certain physical conditions: illness; discomfort; poor hearing, sight, or illumination; a broken pencil; a dislike for the subject, the teacher, or examiner; a failure to give attention to and to comprehend the directions, etc. Any one of these or a dozen other factors may affect the score. Accordingly, there is the possibility and even likelihood of a serious error in the assumption that a test score is a direct evidence of ability. The conditions under which the performance takes place must be known before it is safe to infer ability from performance.

Ability, as an abstract concept, may be defined as the power to do. Power to do, to respond to stimuli and to situations, is the product of training and experience. This suggests that, unless training and native capacity factors are known, inferences as to abilities may be misleading. This point becomes particularly serious in the interpretation of mental-test results, for it is common practice for users of mental tests to infer innate capacity (mental ability) from performance

scores. The real seriousness of this type of uncritical inference may be seen by comparing the interpretations of an achievement-test score and a mental-test score. Both are basically expressions of performance. Equal abilities may be inferred from equal scores from both types of tests if and when all the conditions under which they are given are definitely under control. Although it is difficult to make sure that all physical and physiological factors are adequately controlled in a testing situation, it is possible to regulate most of the mechanical conditions within reasonable limits.

The significant point to note here, however, is the fact that users of achievement tests stop with an inference of equality of ability from equal performance scores, but users of mental tests are obliged to take a further inference. In the interpretation of mental-test results it is common practice to infer equal native capacity from apparent evidences of equal abilities. The fallacies in this argument and the dangers of this step must be readily apparent. *Equal capacities may be inferred from performance scores only when there is direct and positive evidence of two things: first, that the conditions under which the testing took place were identical and equally well controlled; second, that the training opportunities of the individuals compared have been equal.* The mechanics of testing now make it fairly easy to control testing conditions. The second factor represents a real stumbling-block in the way of an accurate and sane interpretation of the mental-test results. The naïve manner in which some makers and many users of mental tests assume equality of learning opportunity, and hence equal capacity from equal performance scores on mental tests, is one of the things which has made teachers and students skeptical of their value.

It is possible that the foregoing discussion of the meaning of a test score may appear to indicate that it is impossible to give meaning to any kind of a test score. Such is not the intention, even though the purpose here is to emphasize the need for a conservative attitude in test-score interpretation. In the long run, the more that is known about the variables underlying test scores, the more critical must the user become. The greatest damage that has been done to the field of educational measurements in the past has come as a direct result of carelessness and ignorance on the part of users of tests, and their tendency to draw unwarranted conclusions from the results. The industrial arts teacher should be able critically to select suitable tests and scales for classroom and shop use, control the mechanical conditions of their administration, and draw sane and defensible conclusions and inferences from the results.

124. Giving Meaning to Informal Test Scores.

The user of industrial education tests in the classroom is confronted with two types of test data for interpretation. The first type, and undoubtedly the more common of the two, deals with the results of informal, teacher-made tests. The results from these home-made tests in turn are of two types: the subjective scores assigned by teachers to pupils' responses to essay-type tests, and the performance scores resulting from informal objective examinations. Although something can be done to improve the interpretation of the relatively unreliable marks assigned to the discussion-type exercise, the performance scores resulting from reasonably long and reliable objective examinations are much more important measures of achievement, and as such deserve complete and accurate interpretation. The second type of educational test data requiring interpretation arises, of course, from the results of using standard tests. Since one of the major functions of the standardization of a test is the establishment of meaning for the test scores, many more types of interpretation are possible for data of this type. Purely for convenience in the organization of this discussion, problems of the interpretation of standard test scores are considered first.

II. NORMS AND STANDARDS

125. The Meaning of Standardization.

Early in the history of objective testing in the classroom practically all that was required for development of a so-called standardized test was to give a few reasonably suitable test exercises to a hundred or more pupils in different school systems. These results were then compiled and submitted as norms. In fact, for many years almost the only real difference between a standardized test and a reasonably good informal objective test was the fact that the former had been tried out in a larger number of different classes. Test standardization as it is now interpreted means much more than the mere derivation of norms, although the existence of norms is still one of the chief distinctive features of the standard tests. There has been much improvement in both the informal test and the more formal standardized test.

In terms of present-day test-construction practices the standardization of a test involves a long period of experimentation with a large body of subject-matter exercises. After the subject-matter field to be tested has been decided upon, there is the very difficult problem of selecting the more important areas of this field to be sampled. Many times experimental evidence must be secured before it is possible to

decide upon the best type of test exercise to use. Even then, many of the exercises prepared in preliminary form for the test are found to be badly stated or to be totally unsuited for the type of test to be constructed. Usually from four to six times as many exercises must be prepared in the preliminary work with a test as will appear in the test itself when in final and standardized form. Special care must be taken to see that a suitable range of difficulty is provided in the items, and that multiple items covering certain of the more important skills are prepared in parallel form so that these items may be adequately sampled in the several forms of the test which must be prepared. After the exercises themselves have been written in preliminary form they must be tried out under experimental conditions in typical classrooms for the purpose of discovering the faulty or ambiguous items and for the additional purpose of discovering the relative difficulty of the several items. From the results of this preliminary use of the exercises two or more roughly scaled forms of the tests may be set up for further experimental use. From the results of this second trial it is usually possible to equate the forms of the tests quite closely by shifting hard and easy items from one form to another until approximate equality is reached. Then the tests are ready for a further trial in a large number of representative classes for the purpose of further equating the forms and establishing norms. It is thus apparent that while standardization is only one of the final steps in the preparation of a carefully made test, it is this extensive sampling of the results of the use of the test in many classrooms which affords the basis for the assignment of meaning to the test scores.

126. Meaning of Norms.

Many of our present-day standardized tests began their existence as informal objective examinations. In fact, many informal examinations of the objective type meet all other criteria of standard tests except that they do not have norms for the evaluation of their scores. Standardized tests are characterized by the fact that they are commonly accompanied by norms representative of the type of accomplishment which may be expected from classes similar to those used in the standardization program. Norms thus furnish the necessary information for the interpretation of the test scores and for the evaluation of achievement of a class. They are obtained by giving the particular test to a large and representative sampling of pupils in the same grades and of a type similar to the group which the teacher wishes to test. To the extent that the sampling is distributed over a large population in typical school situations and the conditions under

which the tests are to be administered are rigidly followed, the norms furnish a reliable and useful basis for interpretation.

127. Standards and Norms.

The use of the term *standardized* in the discussion of tests of the type for which norms are provided has led to the development of a careless tendency to treat the words "standards" and "norms" as being synonymous. The process of securing the data for the critical analysis of tests and the derivation of suitable norms is properly known as *standardizing.* However, *the term standards when used to refer to levels of pupil achievement, implies an ultimate goal to be achieved.* Standards may not actually be reached by any individual, but they are levels of achievement toward which to strive. *Norms are the levels of achievement which typical pupils actually attain.* It is clear that, in the light of these definitions, few tests are accompanied by standards.

128. Specific Uses of Test Norms.

Although the general function of test norms is to provide a basis for the interpretation of test scores, several specific uses should be pointed out at this time. For example, test norms give meaning to the test score. There is no way of determining except through comparison with the norm for a test whether a given score is high or low. To be explicit, is a score of 96 points on the *Newkirk-Stoddard Home Mechanics Test* a high, low, or average score for a pupil to make at the end of the year's course in general shop work? A reference to the norms given in Table 50 will give an answer to this question. As a matter of fact, such a score is so good that only 25 pupils in a hundred may be expected to do better than that.

Norms point out to both pupil and teacher the actual levels or goals of achievement which both should attempt to attain. That is, the norms tell all parties concerned *how far* they have to go and approximately *when* they have arrived. Norms provide almost the only objective basis for the analysis of individual pupil weaknesses. Certain of the better achievement tests in special subject fields are made up of a number of different test-parts designed to measure distinct aspects of achievement in the subject. Many of these tests are provided with separate norms for the special parts of the tests making it possible to reveal the pupil's standing in each of the independent test-parts. (See Table 51.)

Norms for achievement tests used in connection with results from mental tests make it possible to determine within practical limits whether the pupil is working up to the real ability he possesses. The

TABLE 50

END-OF-YEAR NORMS FOR NEWKIRK-STODDARD HOME MECHANICS TEST

	Point Scores	
	Form A	Form B
Mean	74.7	74.6
Median	77.9	73.0
Q_3	95.8	96.3
Q_1	60.5	54.5
S. D.	25.1	27.9

TABLE 51

SCORES ON IOWA SILENT READING TEST: ELEMENTARY, BY A NINTH-GRADE PUPIL

Test	Score	
	Part	Test
1. Paragraph meaning		
A. Science	18	40
B. History	22	
2. Word meaning		
A. General vocabulary	24	42
B. Subject-matter vocabulary	18	
3. Selection of central idea of paragraph	10	10
4. Sentence meaning	28	28
5. Location of information		
A. Alphabetizing	3	12
B. Use of the index	9	
Total comprehension score		132
6. Rate of silent reading		27

basis for this type of analysis of accomplishment is found in the comparison between the mental ability of the pupil expressed in terms of his mental age and his educational achievement as represented by his educational age.

129. Kinds of Norms.

The kind of norm which accompanies a test depends to a large degree upon the level in the school system at which the test is used. The norm is also conditioned somewhat by the nature of the test itself. Tests which are designed for use in the elementary-school grades are usually accompanied by two types of norms, *grade norms, and age norms.* Tests intended for use in the secondary-school grades are usually provided with semester and grade norms only. Age norms do not seem to be particularly useful in the upper grade levels, because so many factors other than age operate to affect achievement. Then, too,

the curve of mental growth flattens out very rapidly in the upper grade levels so that the increments of growth in achievement from age to age at the upper levels are not significant. In place of the age norms for secondary-school and college tests, the common practice today is to provide quite detailed tables of percentile equivalents for the point scores.

In the lower grades, age as well as grade norms are usually provided with achievement tests. In general, the type of norm is determined by the method of grouping the scores when the tabulation for the norms is made. If the pupils are grouped by grades, without respect to age or school progress, the resulting norms are grade norms. If the pupils are classified in accordance with some specific age-scale as the basis for the tabulation, the resulting norms are age norms. In the derivation of grade norms for standard tests it is desirable to have the norms clearly indicate the period they are designed to cover.

III. RESULTS DERIVED FROM NORMS

130. Grade Levels.

Test scores accompanied by a fairly reliable set of grade norms can be expressed in terms of the relative position of these scores with respect to these grade norms. In fact, this is one of the very simple and convenient methods of changing test scores into a form which even the child or his parents can understand. The fact that an individual pupil is in the seventh grade or the eighth grade has come to have some meaning to the average pupil or parent. An isolated test score cannot have this meaning. However, as soon as a test score is identified with a specific grade level of accomplishment it takes on a definite meaning.

TABLE 52

Grade Norms for Haggerty Reading Examination: Sigma 3

Grade	5	6	7	8	9	10	11	12
Score	40	54	68	80	93	104	112	118

The method of deriving the grade levels (so-called G-scores) from grade norms is illustrated from the revised grade norms for the *Haggerty Reading Examination; Sigma 3.* Table 52 shows the scores to be expected at the end of the year for each grade. From this table it is apparent that a score of 93 is the ninth-grade end-of-the-year

norm; a score of 104 is the end of the year norm for the tenth grade, etc. Thus a student making a score of 104 points on this test may be described as achieving at a level equal to an average pupil at the end of the tenth grade. This value may be simply expressed as 10^{10}. A pupil making a score of 93 points on this test may be assigned a grade-level position of 9^{10}, meaning that his achievement is comparable to that expected at the end of the ninth grade. Pupils making test scores between 93 and 104 may be assigned grade levels corresponding to the proportion of the distance between the end of the ninth (beginning tenth) grade and the end of the tenth (beginning eleventh) grade work the scores represent. To illustrate, the score-point distance from 93 to 104 is 11 points. A score of 95 would therefore represent a grade-level distance which is two-elevenths of the way past the beginning of the tenth grade. For practical purposes this is two-tenths of a grade. Thus a score of 95 on this test corresponds roughly to a grade position of 10^2.

131. Age Scores.

Since age equivalents as derived scores have been discussed in this chapter in connection with the meaning of norms, and since they are not considered by most test workers to be of very great significance in the junior-high-school and secondary-school grades they are given no extended treatment here.

132. Percentile Ranks.

One of the favorite ways of interpreting test scores in the secondary-school and college field is to use percentile ranks. Percentile ranks are of particular value in the interpretation of informal and nonstandardized tests, since they permit the comparison of each individual in the group with the group of which he is part. In contrast with the method of assigning ranks by relative position, the calculation of the percentile rank expresses the absolute position of the individual pupil in his relation to the rest of his group. The calculation of *percentiles* involves the division of the total distribution into 100 equal parts, hence the term percentile. Achievement as represented by a test score is expressed as a position in a population of 100 cases. A score representative of high achievement ranks high in the percentile scale and is excelled by only a small number of cases. For example, in Table 53, which shows the percentile norms for a new plane geometry aptitude test, a score of 72 points or more is assigned a percentile score of 100, meaning that such a score is so high that it almost certainly will not be excelled by any one. Table 53 presents the percentile norms for this test in a convenient form for transferring

each possible test score into percentile equivalents. Since percentile scores represent the position of the individual score in a distribution of infinite population, they are very convenient devices for turning test scores from unlike scales into comparable measures. There are many occasions in the interpretation of educational-test results, particularly in experimental situations, when this is very desirable.

TABLE 53

IOWA PLANE GEOMETRY APTITUDE TEST PERCENTILE EQUIVALENTS OF TEST SCORES

$N = 413$ (girls 199; boys 214)

Score	Percentile Equivalents	Score	Percentile Equivalents
72 or more	100	34	55
65 to 72	99	33	53
63 to 64	98	32	50
60	97	31	46
59	96	30	42
58	95	29	40
57	94	28	36
55 to 56	93	27	33
54	92	26	30
53	91	25	27
51 to 52	90	24	25
50	88	23	23
49	87	22	20
48	85	21	18
47	83	20	16
46	81	19	14
45	80	18	12
44	78	17	10
43	76	16	9
42	75	15	8
41	73	14	7
40	70	13	6
39	68	12	5
38	66	11	4
37	63	9 to 10	3
36	60	7 to 8	2
35	58	6	1
		0 to 5	0

133. Intelligence Quotients.

The discovery and use of the concept of *mental age* made possible the development of the quotient idea. In a general way, all quotients derived from results of measurements express the development of the individual as related to average expectancy for his age or mental level.

Scores on mental tests provide the basis for the derivation of mental ages. Scores from achievement tests, provided the tests are accompanied by age norms, may be expressed as *achievement or subject ages*. The ratio between the mental age of an individual student and his chronological age is called an intelligence quotient. If an achievement age is used, the resulting quotient is an educational quotient.

The intelligence quotient (I.Q.) as found in practice is the result of dividing the mental age (M.A.) of the individual by his chronological age (C.A.), both expressed in months. The result of this division is expressed as a whole number by multiplying the quotient by 100. An illustration will make this procedure clear. Let us assume that a pupil who is twelve years and four months of age makes a score on a mental test which gives him a mental-age equivalent of eleven years and three months. At the outset, it is clear that since his mental age is less than his chronological age, he has not made quite normal development in mental ability. That is, his I.Q. will be somewhat less than 100. Actually the I.Q. of this individual is

$$\frac{\text{M. A.}}{\text{C. A.}} = \frac{135}{148} \times 100, \text{ or } 91$$

An intelligence quotient of 100 indicates normal development on the part of the individual. A quotient of less than 100 means that there is more or less retardation in the development, and a quotient of above 100 means more or less accelerated development. It is common practice for examiners in the psychological clinic to consider I.Q.'s of 90 to 110 as average or approximately normal. Quotients above 110 are considered superior in proportion to the extent to which they exceed that value. Similarly, quotients of less than 90 are below average and inferior in proportion to the amount which they fall below that value. I.Q.'s of very high and very low levels are naturally relatively rare. All these interpretations of the intelligence quotient are of course dependent upon the reliability of the measuring instrument on which they are based.

134. Educational Quotients.

Many of the better achievement tests designed for use in the elementary and junior-high-school grades are equipped with age norms which permit the expression of achievement scores as educational ages. These educational-age scores make it possible to derive an educational quotient by following a procedure identical with that used in deriving the intelligence quotient. Since age norms have been found to be impractical for most of the educational achievement tests designed for high school, educational quotients have not been used very widely

in secondary-school measurement. Subjects such as make up the bulk of the industrial arts field do not lend themselves especially well to standardization on an age basis. Hence there is not very great likelihood of using these educational quotients in measurement in industrial education. The same comment appears to hold for the accomplishment quotient (A.Q.), a ratio designed to indicate the relative degree to which an individual student is utilizing his capacity to achieve. The basic idea back of the accomplishment quotient has received some consideration in an earlier section of this chapter.

The various quotients and other measures utilized in the interpretation of educational-test results can be made effective servants of the teacher only through extensive experience in their use. A complete control can be gained only through practice in their calculation and interpretation. Mastery can be retained only through continued use.[1]

IV. RESULTS FROM INFORMAL OBJECTIVE TESTS

135. Objectifying the Marking System.

A critical examination of the marking system and the marks assigned by teachers makes it very apparent that some radical improvements in these phases of educational measurement are needed. As a result of an extensive survey of the problem, and a study of the recommendations of educators who have studied the marking system, the following program for eliminating many of the unsatisfactory features of the present methods of assigning marks is submitted:

1. *Discard the practice of marking pupils in percentages.* Three reasons are advanced for this decision: (*a*) the percentage scale has for its only fixed points 0 and 100. The former means just no ability while the latter means perfect mastery. Yet the complete scale is practically never used in practice. (*b*) The establishment of the limits of the scale fixes the intermediate values. Accordingly, the difference between marks of 75 and 76 should be the same as the difference between marks of 97 and 98. Common observation reveals the absurdity of this assumption. (*c*) The use of the percentage scale presupposes that the teacher is able to distinguish as many as 101 minute differences in accomplishment. Experimental evidence [2] reveals that teachers are able to distinguish from five to seven levels of ability. To use

[1] Extensive opportunity for practice in the derivation of grade equivalents, age scores, percentiles, and quotients of various types is provided in the *Work-Book in Educational Measurements* (Longmans). See particularly Problems 30 to 33.

[2] Ruch, G. M., *The Objective or New-Type Examination,* Scott, Foresman and Company, Chicago, 1929, pages 370-374.

a finer scale assumes an exactness of discrimination on the part of teachers which does not exist. (*d*) The use of an arbitrarily selected percentage as a passing mark as is very commonly done results in throwing the marks into a badly skewed distribution with too large a proportion of the marks piled up at or near the passing mark.

2. *Each mark assigned to a pupil should be a symbol designed to indicate his power to do.* This symbol should be defined in exactly worded statements, understood alike by teachers, administrators, and pupils.

The following definitions of letter grades by Hillbrand[3] are given as an illustration of the type of statements that should be prepared by the teacher for the purpose of defining each of the letter steps in the five-point scale.

Grade	Definition
A	1. Consistently does more than is required.
	2. Has wide vocabulary at his command.
	3. Is always alert; takes active part in discussions.
	4. Has unusual dependability in taking assignments.
	5. Is prompt, neat, and thorough in all work, and unusually free from teachers' correction.
	6. Knows how to select books, tools, materials, and is a rapid worker.
	7. Has initiative and originality in attacking problems.
	8. Has ability to associate and rethink the problem and can adapt himself to new and changing situations.
	9. Has enthusiasm for and interest in his work.
	10. Has ability to apply ideas gained in study to everyday life.
B	1. Frequently does more than is required.
	2. Has good vocabulary and speaks with conviction.
	3. Unusually alive to the situation at hand.
	4. Careful in complying with assignment.
	5. Eager attack on new problems; profits from criticism.
	6. Prompt, neat, thorough, and unusually accurate in all work.
	7. Has ability to apply general principles of the course.
C	1. Does what is required.
	2. Possesses a moderate vocabulary.
	3. Willing to apply himself during class hour.
	4. Does daily preparation with comparative freedom from carelessness.
	5. Attentive to assignments.
	6. Has ability and willingness to comply with instructions and a cheerful response to correction.
	7. Reasonably thorough and prompt in all work.
	8. Has average neatness and accuracy in all work.
	9. Has ability to retain collectively the general principles of the course.

[3] *School and Society,* Vol. 21: 142, January 31, 1925.

GRADE	DEFINITION
D	1. Usually does what is required. 2. Attendance often irregular. 3. Tools and equipment sometimes lacking. 4. Frequently "misunderstands" assignment. 5. Willing but slow in complying with instructions and corrections. 6. Careless in preparation of assignments. 7. Lacking in thoroughness and sometimes tardy with work. 8. Careless in presentation of work.
Fd	1. Usually does a little less than is required. 2. Listless and inattentive in class. 3. Tools and equipment for work often lacking. 4. Always tardy with work. 5. Seldom knows anything outside the lesson. 6. Retains only fragments of the general principles of the course. 7. Lacking in qualities of the first three groups to the extent that he cannot or will not do the work.

3. *Each teacher should give objective examinations or quizzes frequently throughout the term, and the scores from these tests should afford the major basis for his marks.* Prior to the assignment of marks for a school period or semester the pupils should be ranked on their test scores and these scores should then be transformed into marks on a five-point letter scale by the use of the standard deviation technique in large sections or classes (thirty or more). In small classes (less than thirty) this may be accomplished somewhat more simply by dividing the distribution of scores into five groups and assigning the designated mark to previously determined percentages of the class.

The letter grades used and the typical percentages of the class assigned each grade under these conditions are as follows:

Letter Grades	A	B	C	D	Fd
Percentage of class......	4-6	19-21	48-52	19-21	4-6

The essential steps in the assignment of grades by the standard deviation method are outlined in Chapter XIV. The actual solution of a problem utilizing this method in the assignment of marks to objective-test scores from a class of forty-five pupils is shown in Table 45, page 208.

4. *Require teachers to prepare in advance for each six-weeks period carefully worded statements of the objectives of each subject for that period.* Unless this is done, no one can determine whether or not

the pupils are being tested on the points on which they should be tested. This statement of objectives should be the criterion by which the validity of the objective tests is determined.

5. *Work prepared for daily assignments should be treated as a requirement of the course, but marks assigned should be determined by numerous brief objective quizzes or tests upon the work assigned.*

6. *Notebook and laboratory work should be treated as a requirement of the course, and credit should be deducted or withheld for work which is unsatisfactory or incomplete.* However, the marks assigned should be determined by frequent objective tests on the work rather than on the basis of the notebook or laboratory work which may or may not be the pupil's own work.

7. *Assign marks on "accomplishment" or "performance"* rather *than on indefinite subjective factors such as effort, attitude, ability, etc.*

8. *Final grades summarizing all the quiz and test grades for the course can be obtained quite readily by assigning point values to each letter grade, computing the actual average for each pupil, and then re-assigning the class marks on the basis of these averages.* This is a very simple way of assigning final grades for fairly large groups and in courses in which a relatively large number of objective marks are to be summarized in the final grade. It also permits the application of a definite schedule of weighting for certain period and final tests in accordance with the teacher's judgment of their importance.

The accompanying table of point-values (Table 54) corresponding to specific letter grades may be useful to the teacher. Values are suggested for plus and minus values of the letter grades as one means of softening some of the shock from the arbitrariness of letter grades assigned on the basis of the normal curve. Students whose test scores fall just below the point where a superior grade is given sometimes feel that this is a distinct element of unfairness in the system. Assigning the plus and minus letter grades to their quiz scores serves to take care of this problem quite adequately.

V. SUMMARY

This chapter deals with the practical steps in the analysis of test results which make it possible for the classroom teacher to utilize and profit from these results.

The acceptance of the notion that a test score is merely a numerical expression of performance which, subject to the conditions operating at the time, reveals the ability of the individual is essential to a

TABLE 54

SUGGESTED POINT VALUES CORRESPONDING TO LETTER GRADES

Grade	Points
A+	16
A	15
A−	14
B+	12
B	11
B−	10
C+	8
C	7
C−	6
D+	4
D	3
D−	2
Fd	0

safe and sane interpretation of the meaning of test scores. A recognition of the inferences involved in the interpretation of tests of general or mental ability will do much to protect against the over-interpretation of the results of such tests.

The meanings of the terms standardization, standards, and norms are clearly brought out and illustrated in this chapter. The various types of derived scores likely to be useful to the teacher of industrial education are discussed.

Since the major use of tests by the classroom teacher is in the evaluation of achievement, the importance of the informal objective test and other teacher-made measures is emphasized. Present tendencies are distinctly in the direction of the more systematic use of such instruments as the most important single basis for the assignment of teachers' marks. This practical aspect of measurement is so important that considerable attention is given in this chapter to the discussion of possible methods of improving the marking system. After all, the marking system is the one phase of educational measurement with which practically every teacher comes into close contact.

EXERCISES IN INTERPRETING RESULTS OF TESTING

1. Elaborate your interpretation of the basic concepts underlying the meaning of a test score.
2. Show by illustration the real differences between test norms and test standards.
3. By referring to Table 52, compute the grade levels (G-scores) corresponding to scores of 74 and 100.
4. Find the intelligence quotients (I.Q.) for two individuals each 12 years 5 months whose mental ages are 11 years 9 months, and 13 years 11 months, respectively.
5. Criticize the recommendations given for the objectification of the marking system given on pages 235 to 238.
6. In your opinion do the definitions of the meaning of the various letter grades have any place in a program for the objectification of teachers' marks?
7. Show how you would average the following letter grades to secure a term final grade, assuming all grades to count the same except the final examination grade which is allotted triple weight. What final grade would you assign?

First test	— B	Fifth test	— A
Second test	— C	Sixth test	— C
Third test	— B	Final examination	— B
Fourth test	— A		

8. Using the standard deviation technique as illustrated in Table 45, assign letter grades to the objective-test scores secured from the second test given in Problem 7, Chapter XIV.

SELECTED REFERENCES

Bangs, C. W., and Greene, H. A., "Teachers' Marks and the Marking System," *University of Iowa Extension Bulletin*, No. 244, May 15, 1930, Iowa City, Iowa.

Brueckner, Leo J., and Melby, E. O., *Diagnostic and Remedial Teaching* Boston: Houghton Mifflin Company, 1931.

Buckingham, B. R., *Research for Teachers.* New York: Silver, Burdett and Company, 1926.

Franzen, Raymond, *The Accomplishment Quotient Technique.* New York: Teachers College Contributions to Education, No. 125. Columbia University, New York, 1922.

Greene, H. A., *Work-Book in Educational Measurements.* New York: Longmans, Green and Company, 1930.

Greene, H. A., and Jorgensen, A. N., *The Use and Interpretation of Elementary School Tests.* New York: Longmans, Green and Company, 1935.

Odell, C. W., *Educational Measurement in High School.* New York: The Century Company, 1930.

Ruch, G. M., *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Company, 1929.

Rugg, Harold O., "Teachers' Marks and Marking Systems," *Educational Administration and Supervision,* Vol. 1: 117-142, February, 1925.

Smith, H. L., and Wright, W. W., *Tests and Measurements.* New York: Silver, Burdett and Company, 1928.

Trabue, M. R., *Measuring Results in Education.* New York: American Book Company, 1925.

Wilson, G. M., and Hoke, K. J., *How to Measure* (Revised). New York: The Macmillan Company, 1928.

Woody, Clifford, and Sangren, Paul, V., *Administration of the Testing Program.* Yonkers: World Book Company, 1932.

APPENDIX

This appendix contains two types of material supplementary to the discussions and illustrations in the main body of this volume. The list of publishers and distributors of tests should be useful in making a contact with additional types of test materials likely to interest the industrial education teacher. The glossary of terms used in the discussion will help to clarify the meaning of some of the more technical expressions.

APPENDIX A

PRINCIPAL DISTRIBUTORS AND PUBLISHERS OF TESTS OF INTEREST TO INDUSTRIAL EDUCATION TEACHERS AND SUPERVISORS

This Appendix presents a selected list of distributors and publishers of test material likely to be of interest to industrial education students, teachers, and supervisors. Obviously this list does not include many of the important distributors and publishers of tests of more general interest.

Bruce Publishing Company, Milwaukee, Wisconsin.
Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.
Educational Test Bureau, Minneapolis, Minnesota.
Ginn and Company, Boston, Massachusetts.
Gregory Company, The C. A., Cincinnati, Ohio.
Houghton Mifflin Company, Boston, Massachusetts.
Manual Arts Press, Peoria, Illinois.
Marietta Apparatus Company, Marietta, Ohio.
Public School Publishing Company, Bloomington, Illinois.
Scott, Foresman and Company, Chicago, Illinois.
Smith, Turner E., Atlanta, Georgia.
Stanford University Press, Stanford University, California.
Stoelting Company, C. H., Chicago, Illinois.
Teachers College Bureau of Publications, Columbia University, New York.
World Book Company, Yonkers, New York.

APPENDIX B

GLOSSARY

This glossary is appended for the convenience of the student or teacher who may find that many of the terms used in this treatment are outside of his experience.

ability. Power to produce; the result of school training and environment operating on capacity.

accomplishment. Used synonymously with achievement or production.

age norms. Typical performance of subjects grouped by age groups. Usually expressed as the average of actual performance of subjects of different age groups.

age scores. The age equivalents assigned to given point scores on tests provided with age norms.

alternate response. Used in describing any objective test exercise in which the subject must choose between two possible answers, one of which is right and the other of which is wrong. *See* true-false.

ambiguity. A lack of clearness or definiteness in the statement of a fact or a test item.

analytical test. A test which, by taking cross-sections of abilities related to total accomplishment in a subject, furnishes a basis for an analysis of the underlying skills but does not necessarily reveal their interrelationships or causes of weakness.

aptitude. Predisposition for successful achievement in a given field.

arithmetic mean. A measure of central tendency commonly called the average.

array. A collection of data usually arranged around a particular function.

assumed mean. The mid-point of the class interval taken as the zero point in laying off deviations in computing the arithmetic mean from a frequency distribution.

capacity. Power to learn or to profit from training.

character traits. Qualities of the individual such as mentality, honesty, morality, sense of humor, sympathy, etc., which make up personality.

chronological age. The life age of an individual.

classification of pupils. The placement of pupils in a school system in groups by grades or ages for more economical instruction.

coefficient of contingency. A measure of relationship used in the critical analysis of test items. Based upon a comparison of the frequency of cases found in each category with the frequency which we should expect to find if the traits were completely unrelated.

composite score. A single value used to express the results obtained from a number of different measures.

comprehension. The degree of understanding of an exercise or material read.

conditions. Factors causing variations in testing or experimental situations.

correction. A remedy or adjustment. Also in a technical sense in connection with the computation of the arithmetic mean.

correction for chance. In alternate- or multiple-response tests there is a certain opportunity for guessing to enter. The correction for chance is the adjustment for guessing.

corrective. Used synonymously with remedial. Implies the remedying of observed defects or difficulties.

correlation. The relation between two or more series of measures of the same individuals or items.

criterion. The standard by which the validity of measurement may be determined.

diagnosis. Exact identification and location of strengths and weaknesses.

diagnostic test. A test sufficiently reliable and detailed in content to identify and reveal individual pupil weaknesses.

difficulty. When used in reference to test items it implies a large percentage of incorrect responses.

discrimination. The quality in a test or test item which enables it to distinguish adequately between varying levels of ability.

educational guidance. A program designed to direct pupils into school activities in which they are likely to succeed and find most profit, and away from fields in which difficulties and failures are almost certain to be encountered by the child.

error of grouping. A variable error entering into the tabulation of data in frequency distributions. Brought about through the practice of placing together in class intervals measures which may be widely unlike.

error of sampling. The result of using a too limited number of cases as being typical of a large group.

essay-type test. *See* traditional examination.

exercise. A unit of work in a test governed by a specific set of directions. Sometimes used in the sense of a stimulus for drill.

extraversion. The process of being interested in and stimulated by persons and things outside of oneself.

fore exercise. A preliminary or practice exercise for the purpose of giving the pupil experience with the specific test situation.

form. Used to distinguish between two or more closely equivalent arrangements of similar but not identical test items.

frequency. The number of measures in a given interval or tabulation. Frequently indicated by the symbol f.

frequency table. A distribution showing the number of measures assigned to successive class intervals.

fulcrum. The axis upon which a lever is supported and rotated.

general ability. Same as general intelligence. A test of general capacity. Contrasted with achievement.

grade equivalent. The grade or fraction of a grade nearest which a pupil's test score places him when compared with the grade norms for the test.

grouping. The process of classifying data into certain categories.

group test. A test designed for administration to a number of individuals at the same time.

home mechanics. A term used to describe manual tasks arising from the maintenance and repair of household articles and equipment.

individual differences. Observed or measured unlikenesses in pupils in capacity, ability, etc.

informal test. A teacher-made instrument as contrasted with a standardized test.

intelligence. The power to learn, or to profit from training.

I.Q. Intelligence quotient. An index expressing relative brightness as the ratio of mental age to chronological age.

interpolation. The process of locating an intermediate point between two known points in accordance with the operation of laws conditioning the case at hand.

interpretation. The explanation of results and the application of same to a concrete situation.

interval. Used interchangeably with step in preparing a frequency table.

introversion. The process of having one's interests turned in oneself.

manipulative tests. Performance tests in which the subject turns out an objective product as a result of planning and tool operation.

matching-type test. A type of test item in which the stimulus and response forms are presented in parallel columns for convenience in recording the identification.

mean. The arithmetic mean. The point on a scale of values about which the deviations are least.

median. A common measure of central tendency. *See* definition in the text.

mental ability. The power to learn.

mental age. The mental ability of an individual expressed in terms of the age of an average individual having that ability.

mid-point. The exact middle of a step in a frequency table.

multiple-choice test. A type of objective test made up of exercises arranged in such a way that the subject must select one or more correct responses from a group of possible responses.

***N*.** A symbol used to indicate the number of cases in an array.

negative correlation. A relationship in which large values of the one variable are always accompanied by small values in the other.

normal. Typical; making regular progress or development.

norms. Representations of the typical or average performance of subjects of different age or grade groups. Usually based on a large number of cases.

objective. A term used in describing tests in which no opportunity for disagreement as to correctness of response exists.

objectives. Used in the discussion of curriculum construction as synonymous with outcomes.

percentile. The points which divide the total number of cases in a given frequency distribution into 100 equal parts.

performance. Achievement. Also used to distinguish test scores as such from ability or capacity.

personality inventory. A personal rating device from the results of which certain personality characteristics are revealed.

power tests. Tests which express achievement in terms of the difficulty of the task which the subject is just able to perform.

practice effect. Increase in a test score due to previous experience with the test.

product scale. A measuring device listing variable characteristics on which judgments are to be based.

prognostic test. A test designed to predict probable future achievement on the basis of present performance.

quality scale. A device which measures by comparison with a set of standard specimens the result of applying some specific skill.

quartiles. The result of dividing a distribution into quarters.

random sampling. A selection of cases on a purely chance basis.

range. Scale difference between extremes of an array.

rank. Position assigned to a score in a series.

rate test. A device which measures achievement in terms of the number of tasks of uniform difficulty which can be performed in a specified time.

rating scales. Measuring devices which set up levels of qualities or products for the guidance of judges in evaluating such qualities or products in the classroom or shop.

recall test. A test or exercise which calls for the subject to state the answer rather than to recognize it among several possible responses.

recognition test. A test or exercise in which the student merely identifies the correct form of response from several possibilities.

relative rank. Position assigned to scores in a limited array.

reliability. A technical expression of the consistency with which a measuring instrument performs.

reliability coefficient. An index found by the process of correlation indicating the relation which may be expected between successive administrations of the same measuring instrument.

remedial. Material and devices which are designed to correct existing weaknesses in learning or mastery.

retarded. Used to imply school progress or mental or educational development which is slower than is expected of the normal subject.

scores. A description of the performance of a subject.

sigma. Synonymous with standard deviation.

standard deviation. A common measure of variability of scores.

standardization. The process of refining a test and setting up objective goals of performance.

standards. Ultimate goals of achievement. Mistakenly used synonymously with norms which imply actual levels of accomplishment.

subjectivity. The degree to which measurement results are affected by personal factors or judgments.

survey tests. Tests which have for their main purposes the measurement of abilities in terms of broad general functions.

tabulation. The process of classifying data in tables for condensation and interpretation.

teacher's marks. The personal evalution of the pupil's accomplishment in a specific field of activity assigned by the classroom or shop teacher.

technique. Skill in executing tool or machine operations.

test. Any type of measuring device by which a numerical expression of the pupil's performance is secured.

traditional examination. Examinations or tests of the non-objective or discussion type.

training. The learning opportunity afforded through school, shop, or other life contacts.

true-false test. A recognition-type test in which the individual is called upon to determine the truth or falsity of items.

T-scores. A derived test score based on the standard deviation unit.

unit of measurement. The quantity or quality used as the basis for expressing differences.

validity. A term used to express the degree to which a measuring instrument measures the thing it purports to measure.

vocational guidance. A program designed to direct individuals into vocational activities for which they are suited and away from activities for which they are not suited.

zero point of a scale. The point of origin of the instrument.

INDEX